

格致方法·定量研究系列　吴晓刚　主编

社会统计的数学基础

[加] 约翰·福克斯(John Fox) 著
贺光烨 译

★ 革新研究理念

★ 丰富研究工具

★ 最权威、最前沿的定量研究方法指南

1

格致出版社　上海人民出版社



格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用logistic回归分析（第二版）
14. logit与probit：次序模型和多类别模型
15. 定序因变量的logistic回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析
26. 纵贯研究
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图形代数
37. 项目功能差异

上架建议：社会研究方法
ISBN 978-7-5432-2109-3



定价：18.00元
易文网：www.ewen.cc
格致网：www.hibooks.cn

格致方法·定量研究系列 吴晓刚 主编

社会统计的数学基础

[加] 约翰·福克斯(John Fox) 著
贺光烨 译

图书在版编目(CIP)数据

社会统计的数学基础/(加)福克斯(Fox, J.)著;贺光烨译. —上海:格致出版社:上海人民出版社,2012
(格致方法·定量研究系列)
ISBN 978-7-5432-2109-3

Ⅰ. ①社… Ⅱ. ①福… ②贺… Ⅲ. ①社会统计-应用统计学-研究 Ⅳ. ①C91-03 ②O213

中国版本图书馆 CIP 数据核字(2012)第 122666 号

责任编辑 高 璇

格致方法·定量研究系列
社会统计的数学基础
[加]约翰·福克斯 著
贺光烨 译

出 版 世纪出版集团 www.ewen.cc 格致出版社 www.hibooks.cn 上海人民出版社
(200001 上海福建中路193号24层)



编辑部热线 021-63914988
市场部热线 021-63914081

发 行 世纪出版集团发行中心
印 刷 浙江临安曙光印务有限公司
开 本 920×1168 毫米 1/32
印 张 7
字 数 138,000
版 次 2012 年 7 月第 1 版
印 次 2012 年 7 月第 1 次印刷
ISBN 978-7-5432-2109-3/C·64
定 价 18.00 元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书中的35种，翻译成中文，集结成八册，于2011年出版。这八册书分别是:《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》、《数据分析方法五种》和《列表数据分析》。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与SAGE出版社合作，陆续推出新的品种。我们希望本丛书单行本的出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996 年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研

究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院（Institute for Social Research）定量社会研究方法项目的一些讨论会，受益良多。

2003 年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”（Statistics for Social Science）为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课（事实上，大部分硕士生在修完第一门课后都会继续选修第二门课）。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的

方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。“社会科学里的统计学”(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能

力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程，他们目前大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是：

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾"中央"研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约

翰·霍普金斯大学郝令昕教授，也参与了审校工作。

我们希望本丛书的出版，能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

序

曾经有一位社会学的博士研究生跟我说，他要去统计学系上一门基础课程，我问他为什么，他回答："每次在我想更深入地学习高级定量方法时，总感觉有一堵无形的墙。"相对于社会科学院系，统计学系开设的课程更强调数学的基础性，因此，统计学系的学生更容易翻越这堵墙。即便"社会科学的数理基础"这套丛书考虑到了所面对的读者并没有接受足够的数学或统计学训练，然而，近期的许多话题，诸如稳健回归、潜在增长曲线模型等，均需要用到较多更深层次的数学知识，从而使许多读者望而生畏。

《社会统计的数学基础》就是为这些想进一步学习定量方法却时常感到被那堵无形的墙所阻碍的读者而编写的。这本小册子涵盖了许多数学和统计学中容易被人忽视却又至关重要的话题（如矩阵、线性代数、积分、概率理论及统计分布），这些话题经常在统计书籍和论文中出现，许多读者或许以前还接触过，但是对于大多数从事社会科学研究的读者而言，可能还比较陌生。

当得知福克斯的这个项目时，我异常兴奋并积极鼓励他

完成这本书。事实上，许多评论家包括作者本人都感叹，如果类似这样的书可以早出版几年，比如，在他们学习统计的时候，或者在他们为定量方法课程准备授课讲义的时候，那该有多好。

对于这本书，评论家一致认为："它会是协助研究生及社会统计工作者进行研究的得力助手，也会成为大受欢迎的书籍。同时，这本书更将是对定量方法研究的一个重要补充。"

廖福挺

目 录

第 1 章

矩阵、线性代数和几何向量

矩阵为大多数统计提供了一种自然诠释;线性代数是有关线性统计模型的代数计算;几何向量是一种非常强大的概念性工具,它在理解线性代数和标识线性模型等方面很有用。本章的目的是介绍有关矩阵、线性代数和几何向量的基本概念。这些相关话题在社会统计中应用广泛,且其编排形式相对于严格的数学表述来讲是非正式的。一方面,许多计算结果没有提供详尽的根据,而另一方面,这些根据均是提纲挈领的。对更深入的线性代数感兴趣的读者可以参看相关主题的教科书,以获得详细的解释(推荐阅读请参见本章末尾)。

第 1 节 | 矩阵

基本定义

矩阵是一组数字或数字变量的长方形阵列，例如，

$$\underset{(4\times3)}{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix} \qquad [1.1]$$

其更一般地表示为：

$$\underset{(m\times n)}{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad [1.2]$$

像这样的 m 行 n 列矩阵可以称为 m 乘 n 阶矩阵，记做 $(m\times n)$。为方便起见，我有时候在矩阵的下方表示阶，如方程 1.1 和方程 1.2 所表示。矩阵的每一个元或者元素可以用它的行列下标表示，如 a_{ij} 表示矩阵 A 的第 i 行第 j 列元素。若矩阵为单一(实)数，则被称为“纯量”。有时为了简洁方便，我把矩阵中的典型元素放在一个括号里来表示矩阵，如 $\underset{(m\times n)}{A} = \{a_{ij}\}$ 等价于方程 1.2。

一个只有一列元素的矩阵为列向量，如

$$\underset{(m\times 1)}{\mathbf{a}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

与之类似，一个只有一行元素的矩阵为行向量，

$$\mathbf{b}' = [b_1, \quad b_2, \quad \cdots, \quad b_n]$$

为了便于区分，我在行向量的元素间加上了逗号。

矩阵 **A** 的转置表示为 **A**′，它是将 **A** 的第 i 行转变为 **A**′ 的第 i 列所构成，因此以方程 1.1 和方程 1.2 为基准，则有：

$$\underset{(3\times 4)}{\mathbf{X}'} = \begin{bmatrix} 1 & 4 & 7 & 0 \\ -2 & -5 & 8 & 0 \\ 3 & -6 & 9 & 10 \end{bmatrix}$$

$$\underset{(n\times m)}{\mathbf{A}'} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

请注意，$(\mathbf{A}')' = \mathbf{A}$。通常，我所说的向量是指列向量（如上面的 **a**），除非明确指出它是被转置的（如 **b**′）。

N 阶矩阵，正如它的名字一样，拥有 n 行 n 列。元素 a_{ii}（例如，a_{11}，a_{22}，…，a_{nn}）组成了方阵 **A** 的主对角线。对角线上所有元素的和叫做矩阵的“迹”：

$$\text{trace}(A) \equiv \sum_{i=1}^{n} a_{ii}$$

如方阵

$$\underset{(3\times3)}{\mathbf{B}} = \begin{bmatrix} -5 & 1 & 3 \\ 2 & 2 & 6 \\ 7 & 3 & -4 \end{bmatrix}$$

其对角线元素为−5、2和−4，因此迹为 $\sum_{i=1}^{3} b_{ii} = -5+2-4 = -7$。

如果 $\mathbf{A} = \mathbf{A}'$，则称该方阵是对称的，即对于所有的 i 和 j，$a_{ij} = a_{ji}$。根据定义可知，(上面的)方阵 $\mathbf{B}$ 是不对称的，而方阵

$$\mathbf{C} = \begin{bmatrix} -5 & 1 & 3 \\ 1 & 2 & 6 \\ 3 & 6 & -4 \end{bmatrix}$$

是对称的。统计应用中的许多矩阵都是对称的，如相关性矩阵、协方差矩阵、平方和矩阵或者交叉乘积矩阵。

上三角矩阵是指主对角线下方的元素都为0的矩阵：

$$\underset{(n\times n)}{\mathbf{U}} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

下三角矩阵指主对角线上方的元素都为0的矩阵：

$$\underset{(n\times n)}{\mathbf{L}} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$$

对角矩阵是指除主对角线外，其他元素都为 0 的矩阵：

$$\underset{(n\times n)}{\mathbf{D}} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}$$

为简单起见，我将对角矩阵 $\mathbf{D}$ 表示为 $\mathbf{D} = \text{diag}(d_1, d_2, \cdots, d_n)$。纯量矩阵是所有元素都相等的对角矩阵：$\mathbf{S} = \text{diag}(s_1, s_2, \cdots, s_n)$。一种重要的纯量矩阵是单位矩阵，它的主对角线上的元素全是 1：

$$\underset{(n\times n)}{\mathbf{I}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

我一般将 $\underset{(n\times n)}{\mathbf{I}}$ 写成 $\mathbf{I}_n$。

另外两种重要的纯量矩阵是零矩阵（所以元素都为 0）和向量 1（所有元素都为 1）。我用 $\mathbf{1}_n$ 表示 n 元向量，如 $\mathbf{1}_4 = [1, 1, 1, 1]'$。尽管单位矩阵、零矩阵和向量 1 都属于矩阵，但是为方便起见，我们通常指定它们为奇异矩阵，如单位矩阵就是一个奇异矩阵。

分块矩阵是指将一个矩阵的元素分归于若干子矩阵，如

$$\underset{(4\times 3)}{\mathbf{A}} = \left[\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ \hline a_{41} & a_{42} & a_{43} \end{array}\right] = \left[\begin{array}{c|c} \underset{(3\times 2)}{\mathbf{A}_{11}} & \underset{(3\times 1)}{\mathbf{A}_{12}} \\ \hline \underset{(1\times 2)}{\mathbf{A}_{21}} & \underset{(1\times 1)}{\mathbf{A}_{22}} \end{array}\right]$$

其中，子矩阵 $\mathbf{A}_{11}$ 为：

$$\mathbf{A}_{11} \equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

同样，$\mathbf{A}_{12}$、$\mathbf{A}_{21}$、$\mathbf{A}_{22}$具有类似的定义。当没有歧义时，我会将子矩阵间的线省略。如果一个矩阵仅被垂直分割，我会用逗号来区分子矩阵，例如，

$$\underset{(m\times n+p)}{\mathbf{C}} = \left[\underset{(m\times n)}{\mathbf{C}_1}, \underset{(m\times p)}{\mathbf{C}_2}\right]$$

简单矩阵运算法则

如果两个矩阵具有相同的阶且它们相应的元素都相等，那么，我们说这两个矩阵相等。

当且仅当两个矩阵同阶时，它们才可以相加，通过将两个矩阵中的对应元素相加，即可得到矩阵的和。因此，当 $\mathbf{A}$ 和 $\mathbf{B}$ 均为 $(m\times n)$ 阶时，那么，$\mathbf{C}=\mathbf{A}+\mathbf{B}$，其阶仍为 $(m\times n)$，且 $c_{ij}=a_{ij}+b_{ij}$。同样，如果 $\mathbf{D}=\mathbf{A}-\mathbf{B}$，那么，$\mathbf{D}$ 的阶也为 $(m\times n)$，且 $d_{ij}=a_{ij}-b_{ij}$。如果要求矩阵 $\mathbf{A}$ 的负矩阵 $\mathbf{E}$，即 $\mathbf{E}=-\mathbf{A}$，它的阶同 $\mathbf{A}$ 相等，则 $e_{ij}=-a_{ij}$。例如：

$$\underset{(2\times 3)}{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

和

$$\underset{(2\times 3)}{\mathbf{B}} = \begin{bmatrix} -5 & 1 & 2 \\ 3 & 0 & -4 \end{bmatrix}$$

我们得到：

$$\underset{(2\times3)}{\mathbf{C}} = \mathbf{A}+\mathbf{B} = \begin{bmatrix} -4 & 3 & 5 \\ 7 & 5 & 2 \end{bmatrix}$$

$$\underset{(2\times3)}{\mathbf{D}} = \mathbf{A}-\mathbf{B} = \begin{bmatrix} 6 & 1 & 1 \\ 1 & 5 & 10 \end{bmatrix}$$

$$\underset{(2\times3)}{\mathbf{E}} = -\mathbf{B} = \begin{bmatrix} 5 & -1 & -2 \\ -3 & 0 & 4 \end{bmatrix}$$

由于这些计算均是针对元素的运算，所以矩阵相加、相减及求其负矩阵的依据都与纯量运算法则相同。特别是：

$$\mathbf{A}+\mathbf{B} = \mathbf{B}+\mathbf{A}\text{（矩阵相加的交换律）}$$

$$\mathbf{A}+(\mathbf{B}+\mathbf{C}) = (\mathbf{A}+\mathbf{B})+\mathbf{C}\text{（矩阵相加的结合律）}$$

$$\mathbf{A}-\mathbf{B} = \mathbf{A}+(-\mathbf{B}) = -(\mathbf{B}-\mathbf{A})$$

$$\mathbf{A}-\mathbf{A} = \mathbf{0}$$

$$\mathbf{A}-\mathbf{0} = \mathbf{A}$$

$$-(-\mathbf{A}) = \mathbf{A}$$

$$(\mathbf{A}+\mathbf{B})' = \mathbf{A}'+\mathbf{B}'$$

一个 $(m\times n)$ 阶的矩阵 $\mathbf{A}$ 与一个纯量 c 的乘积为 $\mathbf{B} = c\mathbf{A}$，其中，$b_{ij} = ca_{ij}$。续前例，我们得到：

$$\underset{(2\times3)}{\mathbf{F}} = 3\times\mathbf{B} = \mathbf{B}\times3 = \begin{bmatrix} -15 & 3 & 6 \\ 9 & 0 & -12 \end{bmatrix}$$

纯量与矩阵的乘积遵循如下法则：

$$c\mathbf{A} = \mathbf{A}c\text{（交换律）}$$

$$\mathbf{A}(b+c) = \mathbf{A}b+\mathbf{A}c\text{（纯量分配律）}$$

$$c(\mathbf{A}+\mathbf{B}) = c\mathbf{A}+c\mathbf{B}\text{（矩阵分配律）}$$

$$0\mathbf{A} = \mathbf{0}$$

$$1\mathbf{A} = \mathbf{A}$$

$$(-1)\mathbf{A} = -\mathbf{A}$$

其中，b、c、0、1 和 -1 都是纯量，**A**、**B** 和 **0** 为同阶矩阵。

两个 n 元向量的内积（或者点乘）为一个纯量，它是通过相加相对应向量元的乘积得来的。

$$\mathbf{a}' \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

例如，

$$[2,\ \ 0,\ \ 1,\ \ 3] \cdot \begin{bmatrix} -1 \\ 6 \\ 0 \\ 9 \end{bmatrix} = 2(-1)+0(6)+1(0)+3(9)$$

$$= 25$$

当矩阵 **A** 的列数与矩阵 **B** 的行数相等时，我们说矩阵 **A** 和矩阵 **B** 是乘法相适的。因此，当 **A** 为 $(m\times n)$ 阶，**B** 为 $(n\times p)$ 阶时，矩阵 **A** 和 **B** 乘法相适（如下例）。

$$\underset{(2\times3)}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}} \underset{(3\times3)}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}$$

但是以下矩阵却不乘法相适：

$$\underset{(3\times3)}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \underset{(2\times3)}{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}}$$

让 **C** = **AB** 作为矩阵 **A**、矩阵 **B** 的乘积；让 $\mathbf{a}_i$ 代表 **A** 第 i 行，$\mathbf{b}_j$ 代表 **B** 第 j 列，那么，我们知道，**C** 就是一个 $(m\times p)$ 的矩阵，且 $c_{ij} = \mathbf{a}_i' \cdot \mathbf{b}_j = \sum_{k=1}^{n} a_{ik} b_{kj}$。

请看下面几个例子：

$$\underset{(2\times3)}{\begin{bmatrix} \Rightarrow & & \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}}\underset{(3\times3)}{\begin{bmatrix} \Downarrow & 1 & 0 & 0 \\ & 0 & 1 & 0 \\ & 0 & 0 & 1 \end{bmatrix}}$$

$$=\underset{(2\times3)}{\begin{bmatrix} 1(1)+2(0)+3(0), & 1(0)+2(1)+3(0), & 1(0)+2(0)+3(1) \\ 4(1)+5(0)+6(0), & 4(0)+5(1)+6(0), & 4(0)+5(0)+6(1) \end{bmatrix}}$$

$$=\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

$$\underset{(1\times4)}{[\beta_0, \quad \beta_1, \quad \beta_2, \quad \beta_3]}\underset{(4\times1)}{\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}}=\underset{(1\times1)}{[\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3]}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix}=\begin{bmatrix} 4 & 5 \\ 8 & 13 \end{bmatrix} \qquad [1.3]$$

$$\begin{bmatrix} 0 & 3 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}=\begin{bmatrix} 9 & 12 \\ 5 & 8 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad [1.4]$$

$$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

在第一个例子中，箭头表示左边矩阵里的元素如何与右边矩

阵里的元素相乘。

矩阵乘法遵循结合律，$\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$，其分配律同加法相似：

$$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$$

$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$$

但是它又不是广义上的可交换：如果 $\mathbf{A}$ 为 $(m \times n)$ 阶，$\mathbf{B}$ 为 $(n \times p)$ 阶，矩阵 $\mathbf{AB}$ 如之前所定义的是乘法相适的，但是 $\mathbf{BA}$ 要乘法相适必须满足 $m = p$ 。即便满足了这个条件，$\mathbf{AB}$ 和 $\mathbf{BA}$ 的阶也可能不同。而且即使 $\mathbf{A}$ 和 $\mathbf{B}$ 同阶且都为 (2×2)，即乘积 $\mathbf{AB}$ 和 $\mathbf{BA}$ 的阶也相同，但是所得矩阵仍然不同，这点可参见方程 1.3。除非 $\mathbf{A}$ 与 $\mathbf{B}$ 如方程 1.4 所示，$\mathbf{AB}=\mathbf{BA}$，我们可以说，$\mathbf{A}$ 与 $\mathbf{B}$ 的乘积满足交换律，否则轻易下结论说 $\mathbf{AB} = \mathbf{BA}$ 是错误的。然而，纯量可以在矩阵乘积中随意摆放而不影响计算结果：$c\mathbf{AB} = \mathbf{A}c\mathbf{B} = \mathbf{AB}c$。

单位矩阵和零矩阵在矩阵的乘法中扮演着非常重要的角色，因为它与含有数字 0 和 1 的纯量运算相似。

$$\underset{(m\times n)}{\mathbf{A}}\mathbf{I}_n = \mathbf{I}_m\underset{(m\times n)}{\mathbf{A}} = \mathbf{A}$$

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times p)}{\mathbf{0}} = \underset{(m\times p)}{\mathbf{0}}$$

$$\underset{(q\times m)}{\mathbf{0}}\ \underset{(m\times n)}{\mathbf{A}} = \underset{(q\times n)}{\mathbf{0}}$$

矩阵乘积还有一个性质在纯量运算中没有，即 $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$，两矩阵之积的转置是它们顺序相反的转置矩阵之积。这可推广为：

$$(\mathbf{AB}\cdots\mathbf{F})' = \mathbf{F}'\cdots\mathbf{B}'\mathbf{A}'$$

一个矩阵的平方是它和它本身的乘积，即 $\mathbf{A}^2 = \mathbf{AA}$，

$\mathbf{A}^3=\mathbf{AAA}=\mathbf{AA}^2=\mathbf{A}^2\mathbf{A}$，以此类推。如果 $\mathbf{B}^2=\mathbf{A}$，那么我们就可以说 $\mathbf{B}$ 是 $\mathbf{A}$ 的平方根，或者我们可以将 $\mathbf{B}$ 写成 $\mathbf{A}^{1/2}$。与纯量计算不同，一个矩阵的平方根不是唯一的，当然，纯量的平方根也不是唯一的，但区别仅在于符号。如果 $\mathbf{A}^2=\mathbf{A}$，那么，我们称 $\mathbf{A}$ 为"等幂元"。在纯量运算中，依照惯例，$\mathbf{A}^0=\mathbf{I}$（其中，$\mathbf{I}$ 与 $\mathbf{A}$ 同阶）。矩阵 $\mathbf{A}$ 的逆矩阵记为 $\mathbf{A}^{-1}$，其矩阵元素并不等于 $\{1/a_{ij}\}$。

为了便于讲解矩阵的加法、减法及乘法，我们常常把分块矩阵的子矩阵看做矩阵里的元素，只要这些元素分割恰当。例如，

$$\mathbf{A}=\left[\begin{array}{ccc|cc} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{array}\right]=\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

并且，

$$\mathbf{B}=\left[\begin{array}{ccc|cc} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} \\ \hline b_{31} & b_{32} & b_{33} & b_{34} & b_{35} \end{array}\right]=\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$$

那么，

$$\mathbf{A}+\mathbf{B}=\left[\begin{array}{c|c} \mathbf{A}_{11}+\mathbf{B}_{11} & \mathbf{A}_{12}+\mathbf{B}_{12} \\ \hline \mathbf{A}_{21}+\mathbf{B}_{21} & \mathbf{A}_{22}+\mathbf{B}_{22} \end{array}\right]$$

同样，如果

$$\underset{(m+n\times p+q)}{\mathbf{A}}=\begin{bmatrix} \underset{(m\times p)}{\mathbf{A}_{11}} & \underset{(m\times q)}{\mathbf{A}_{12}} \\ \underset{(n\times p)}{\mathbf{A}_{21}} & \underset{(n\times q)}{\mathbf{A}_{22}} \end{bmatrix}$$

$$\underset{(p+q\times r+s)}{\mathbf{B}} = \begin{bmatrix} \underset{(p\times r)}{\mathbf{B}_{11}} & \underset{(p\times s)}{\mathbf{B}_{12}} \\ \underset{(q\times r)}{\mathbf{B}_{21}} & \underset{(q\times s)}{\mathbf{B}_{22}} \end{bmatrix}$$

那么，

$$\underset{(m+n\times r+s)}{\mathbf{AB}} = \left[\begin{array}{c|c} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \hline \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{array}\right]$$

根据矩阵的定义，我们可以把纯量方程组用矩阵方程来表示。考虑下面含有两个未知变量(x_1，x_2)的线性方程组：

$$2x_1 + 5x_2 = 4$$

$$x_1 + 3x_2 = 5$$

这些方程之所以为线性，是因为其相加之和为常数(如第一个方程的右边)，而且方程左边均是常数和一次变量的乘积(如第一个方程左边的第一项 $2x_1$)。

$2x_1 + 5x_2 = 4$ 和 $x_1 + 3x_2 = 5$ 这两个方程分别代表一个二维坐标空间。我们可以把以上方程组用矩阵方程来表示，得到：

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

$$\underset{(2\times 2)}{\mathbf{A}}\ \underset{(2\times 1)}{\mathbf{x}} = \underset{(2\times 1)}{\mathbf{b}}$$

其中，

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

有关线性方程组的组成和解法会在后文中给予详解。

逆矩阵

在纯量计算中，除法是解简单方程的重要工具，例如，

$$6x = 12$$

$$x = \frac{12}{6} = 2$$

或者，

$$\frac{1}{6} \times 6x = \frac{1}{6} \times 12$$

$$x = 2$$

其中，$\frac{1}{6} = 6^{-1}$，即纯量 6 的倒数。

在矩阵计算中，没有直接的除法，但是大多数方形矩阵都有逆矩阵。一个方形矩阵的逆矩阵[1]是一个同阶的方形矩阵，记做 $\mathbf{A}^{-1}$。它有如下性质：$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$。如果方形矩阵可逆，则称其为“非奇异矩阵”（当数学家第一次遇到非零且不可逆的矩阵时，他们发现这种矩阵存在的数量显著，因此称这种性质为“奇异性”）。如果一个矩阵存在逆矩阵，那么它就具有唯一性。对于一个方形矩阵 $\mathbf{A}$，$\mathbf{AB} = \mathbf{I}$，那么必然有 $\mathbf{BA} = \mathbf{I}$，因此 $\mathbf{B} = \mathbf{A}^{-1}$。请看一个非奇异矩阵：

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}$$

它的逆矩阵为：

$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}$$

我们可以证明：

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 3 & -5 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

在纯量代数中，只有 0 没有倒数。我们接下来举一个关于非零奇异矩阵的列子，假设 **B** 为矩阵 **A** 的逆矩阵，

$$\mathbf{A}=\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

但是，

$$\mathbf{AB}=\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}=\begin{bmatrix} b_{11} & b_{12} \\ 0 & 0 \end{bmatrix}\neq \mathbf{I}_2$$

这与假设相悖，因此，我们说 **A** 没有逆矩阵。

寻找非奇异方形矩阵的逆矩阵有很多方法，在这里，我来简单介绍一种方法——高斯消去法。尽管在计算机执行时许多方法都可以提供精确的结果，但是消去法使用起来较简单，且在应用范围上也超出了矩阵求逆（这一点我们在后面的内容中会有所提及）。现在我们以如下矩阵为例：

$$\begin{bmatrix} 2 & -2 & 0 \\ 1 & -1 & 1 \\ 4 & 4 & -4 \end{bmatrix} \qquad [1.5]$$

首先，把该矩阵与单位矩阵合并，即构造一个分块或者增广矩阵：

$$\left[\begin{array}{ccc|ccc} 2 & -2 & 0 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$

然后，我们试图通过以下操作把原来的矩阵变为单位矩阵：

E_I：用一个非零纯量与矩阵的任意一行相乘。

E_{II}：把矩阵中某一行的倍数加到另一行上。

E_{III}：交换两行。

E_I、E_{II}和E_{III}被称为“初等行变换”。

从第一行开始，我们对每一行轮流进行初等行变换，同时保证对角线上的元素不能为0，如果遇到对角线元素为0的情况，我们可以把对应的那一行移到下一行，然后用行元素除以这一行的对角线元素（该元素也称为“主元”）。最后，用这一行的倍数加上另外的行，以消除其他行对角线元素以外的非零元素。具体过程如下所示：

1. 增广矩阵第一行除以2，

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 1 & 0 & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$

2. 第二行减去第一行，

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \\ 4 & 4 & -4 & 0 & 0 & 1 \end{array}\right]$$

3. 用第三行减去第一行乘以 4，

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \\ 0 & 8 & -4 & -2 & 0 & 1 \end{array}\right]$$

4. 由于第二行对角线元素为 0，所以将第二行、第三行交换，

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 8 & -4 & -2 & 0 & 1 \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

5. 第二行除以 8，

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

6. 第一行加第二行，

$$\left[\begin{array}{ccc|ccc} 1 & 0 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

7. 因为第一行主元已经为 1，所以用第三行乘以$\frac{1}{2}$再加第一行，

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

8. 第三行乘以$\frac{1}{2}$再加第二行，

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{8} \\ 0 & 1 & 0 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{2} & 1 & 0 \end{array}\right]$$

当原先的矩阵变为单位矩阵时，增广矩阵的最后三列则包含原先矩阵之逆阵，我们可以通过以下步骤来证明：

$$\begin{bmatrix} 2 & -2 & 0 \\ 1 & -1 & 1 \\ 4 & 4 & -4 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{8} \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{8} \\ -\frac{1}{2} & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \checkmark$$

解释消去法可行性的方法很简单：每个初等行变换都可以用一个矩阵乘法来表示。因此，当我们要交换第二行和第三行时，我们只需在原矩阵的左边乘以以下矩阵：

$$\mathbf{E}_{III} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

步骤中包含一系列 p 个对增广矩阵 $[\underset{(n\times n)}{\mathbf{A}}, \mathbf{I}_n]$ 的初等行变换，我们可以写成：

$$\mathbf{E}_p\cdots\mathbf{E}_2\mathbf{E}_1[\mathbf{A}, \mathbf{I}_n] = [\mathbf{I}_n, \mathbf{B}]$$

其中，$\mathbf{E}_i$ 表示第 i 个变换。定义 $\mathbf{E} \equiv \mathbf{E}_p\cdots\mathbf{E}_2\mathbf{E}_1$，即 $\mathbf{EA} = \mathbf{I}_n$（暗示了 $\mathbf{E} = \mathbf{A}^{-1}$），$\mathbf{EI}_n = \mathbf{B}$。因此，$\mathbf{B} = \mathbf{E} = \mathbf{A}^{-1}$。如果 $\mathbf{A}$ 为奇异矩阵，那么，它就无法通过初等行变换转为单位矩阵 $\mathbf{I}$。在该过程中，非零主元不存在。

矩阵逆阵遵循以下法则：

$$\mathbf{I}^{-1} = \mathbf{I}$$

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$$

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

$$(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$$

其中，$\mathbf{A}$ 和 $\mathbf{B}$ 为 n 阶非奇异矩阵，c 为非零纯量。如果 $\mathbf{D} = \mathrm{diag}(d_1, d_2, \cdots, d_n)$，且所有 $d_i \neq 0$，那么 $\mathbf{D}$ 是一个非奇异矩阵，$\mathbf{D}^{-1} = \mathrm{diag}\left(\frac{1}{d_1}, \frac{1}{d_2}, \cdots, \frac{1}{d_n}\right)$。最后，一个非奇异对称矩阵的逆矩阵也是对称的。

行列式

对应于每个方形矩阵 **A**,都有一个称为“矩阵行列式”的数,这个数记做det **A**。[2]对于一个 (2×2) 的矩阵,其行列式可表示为 $\det \mathbf{A}=a_{11}a_{22}-a_{12}a_{21}$。对于一个 (3×3) 的矩阵,其行列式可表示为:

$$\det \mathbf{A}=a_{11}a_{22}a_{33}-a_{11}a_{23}a_{32}+a_{12}a_{23}a_{31}-a_{12}a_{21}a_{33}+a_{13}a_{21}a_{32}-a_{13}a_{22}a_{31}$$

尽管对于 n 阶方形矩阵的行列式有一个广义的定义,但是我认为,用以下性质(或定理)来描述行列式更简单:

D1:如果用纯量常数乘以矩阵 **A** 的某一行,那么矩阵的新行列式则为原来行列式与该常数的乘积。

D2:如果把矩阵 **A** 某一行的若干倍加到另一行,行列式值不变。

D3:交换矩阵 **A** 的任意两行会改变行列式的符号。

D4:$\det \mathbf{I}=1$

定理 D1、D2 和 D3 指出了三种初等行变换对行列式的影响。由于上述高斯消去法可将一个方形矩阵转变为单位矩阵,因此,这些性质加上定理 D4 已经可以充分确定行列式的值。行列式可以简单地通过主元乘积得到,在消去过程中,如果使用了一次偶数行交换,则要在乘积前面加负号。如方程 1.5,其行列式等于 $-(2)(8)(1)=-16$,因为在第四步有一

个行交换(第二行和第三行),通过步骤1、步骤5及步骤7,我们知道矩阵主元分别为2、8和1。如果矩阵为奇异矩阵,那么则有一个或者一个以上的主元为0,因此行列式为0。相反,对于一个非奇异矩阵,其主元不可能为0。

行列式有时会在统计应用中直接出现,例如,出现在多元正态分布的公式中。

克罗内克积

假设**A**是一个$m\times n$阶矩阵,**B**为一个$p\times q$阶矩阵。那么**A**和**B**的克罗内克积记做**A**⊗**B**,定义为:

$$\underset{(mp\times nq)}{\mathbf{A}\otimes\mathbf{B}}=\begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

由于克罗内克积可以表示分块矩阵,因此在统计中非常有用。例如,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\otimes\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}=\left[\begin{array}{cc|cc|cc} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \sigma_1^2 & \sigma_{12} & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_2^2 \end{array}\right]$$

克罗内克积的许多性质与普通矩阵相似,尤其是,

$$\mathbf{A}\otimes(\mathbf{B}+\mathbf{C})=\mathbf{A}\otimes\mathbf{B}+\mathbf{A}\otimes\mathbf{C}$$
$$(\mathbf{B}+\mathbf{C})\otimes\mathbf{A}=\mathbf{B}\otimes\mathbf{A}+\mathbf{C}\otimes\mathbf{A}$$
$$(\mathbf{A}\otimes\mathbf{B})\otimes\mathbf{D}=\mathbf{A}\otimes(\mathbf{B}\otimes\mathbf{D})$$
$$c(\mathbf{A}\otimes\mathbf{B})=(c\mathbf{A})\otimes\mathbf{B}=\mathbf{A}\otimes(c\mathbf{B})$$

其中,**B** 和 **C** 为同阶矩阵,c 为纯量。如同矩阵乘法,克罗内克积不具交换性,从广义上说,$\mathbf{A}\otimes\mathbf{B}\neq\mathbf{B}\otimes\mathbf{A}$。另外,对于矩阵 $\underset{(m\times n)}{\mathbf{A}}$、$\underset{(p\times q)}{\mathbf{B}}$、$\underset{(n\times r)}{\mathbf{C}}$ 和 $\underset{(q\times s)}{\mathbf{D}}$,

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D})=\mathbf{AC}\otimes\mathbf{BD}$$

如果 $\underset{(n\times n)}{\mathbf{A}}\underset{(m\times m)}{\mathbf{B}}$ 为非奇异矩阵,那么,

$$(\mathbf{A}\otimes\mathbf{B})^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$$

因为,

$$\begin{aligned}(\mathbf{A}\otimes\mathbf{B})(\mathbf{A}^{-1}\otimes\mathbf{B}^{-1})&=(\mathbf{AA}^{-1})\otimes(\mathbf{BB}^{-1})\\&=\mathbf{I}_n\otimes\mathbf{I}_m=\mathbf{I}_{(nm\times nm)}\end{aligned}$$

最后,对于任意矩阵 **A** 和 **B**,

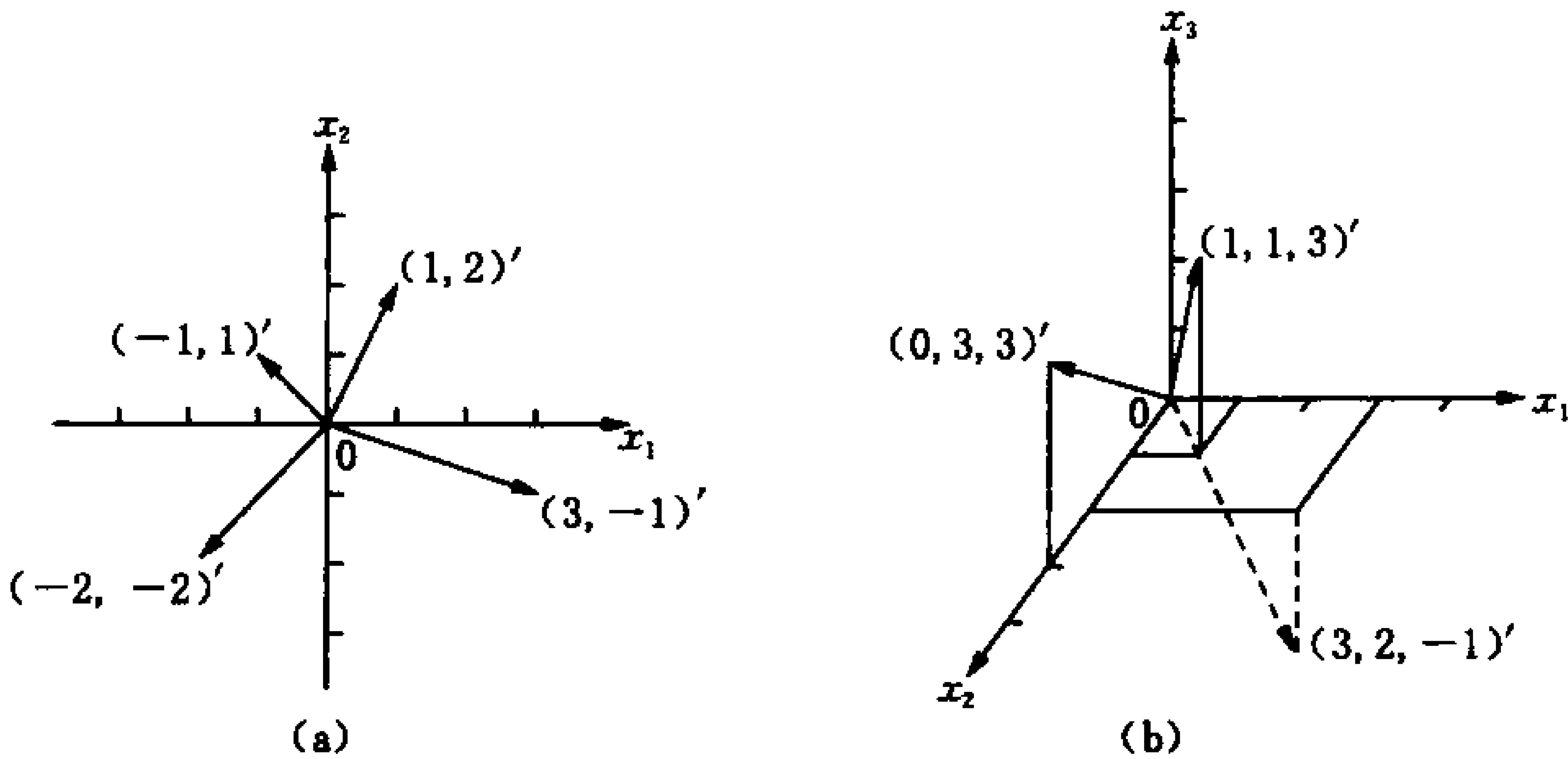
$$(\mathbf{A}\otimes\mathbf{B})'=\mathbf{A}'\otimes\mathbf{B}'$$

对于分别具有 m 和 n 阶的方形矩阵 **A** 和 **B**,

$$\text{trace}(\mathbf{A}\otimes\mathbf{B})=\text{trace}(\mathbf{A})\times\text{trace}(\mathbf{B})$$
$$\det(\mathbf{A}\otimes\mathbf{B})=(\det\mathbf{A})^m(\det\mathbf{B})^n$$

第 2 节 | 基础几何向量

在代数中，向量为只含有一列(或者一行)的矩阵。其几何解释为:向量 $\mathbf{x}=[x_1, x_2, \cdots, x_n]$，表示 n 维笛卡尔坐标空间中的零点到(由向量元决定的)终点的有向线段。有关二维或三维空间的向量例子，请参见图 1.1。

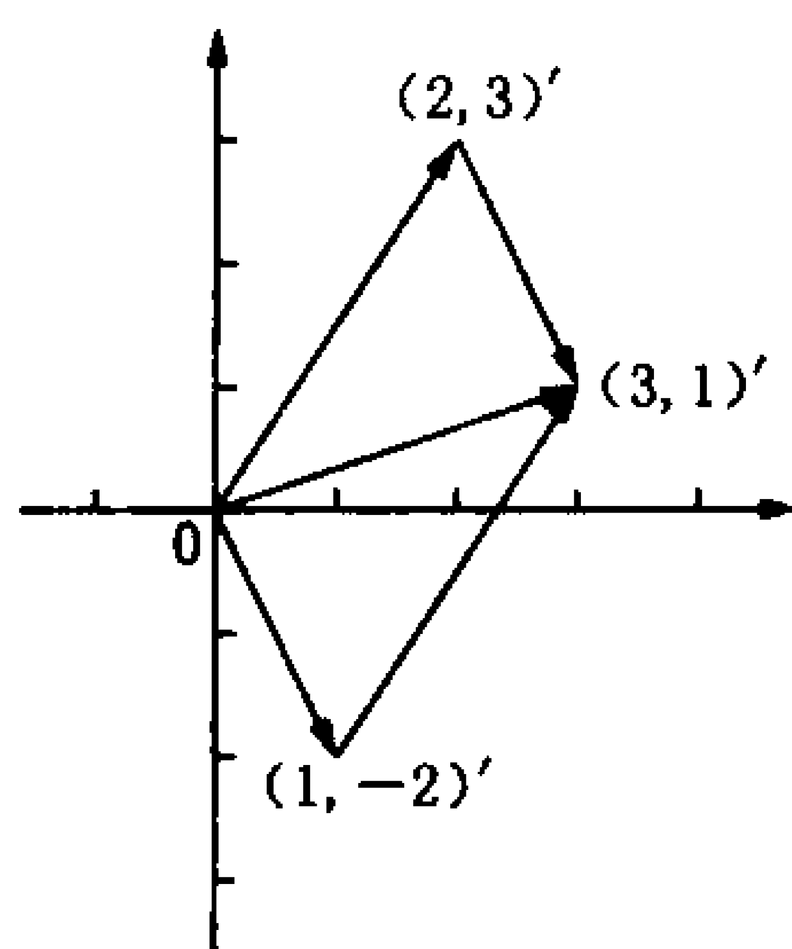


注:每个向量都是一个以 0 为起点的有向线段，其终点由向量元决定。

图 1.1　几何向量举例:(a)二维平面;(b)三维空间

有关向量基本算术的几何说明非常简单，已知长度和方向，我们就可以确定一个向量，不论其起点是不是在坐标零点。两个向量 $\mathbf{x}_1$、$\mathbf{x}_2$ 相加，只要使其中一个向量 $\mathbf{x}_1$ 平移到其终点与另一个向量 $\mathbf{x}_2$ 的起点重合，此时所得的以 $\mathbf{x}_1$ 的起点为起点、以 $\mathbf{x}_2$ 的终点为终点的向量即由加法所得的向量，同

时，该平移向量的长度与方向（与所有坐标轴所成的角度）保持不变。图 1.2在二维坐标系里描述了向量加法的操作。它等同于以 $\mathbf{x}_1$、$\mathbf{x}_2$ 为邻边作平行四边形，以坐标 0 点为起点的对角线即向量 $\mathbf{x}_1$、$\mathbf{x}_2$ 的和。



注：把其中一个向量平移到其终点与另一个向量的起点相重合构成一个平行四边形，以坐标 0 点为起点的对角线即两向量的和。

图 1.2 两个向量相加

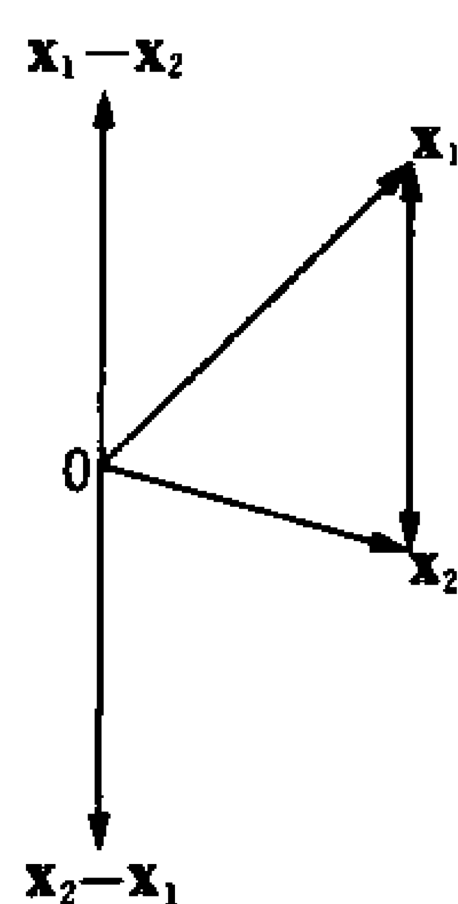


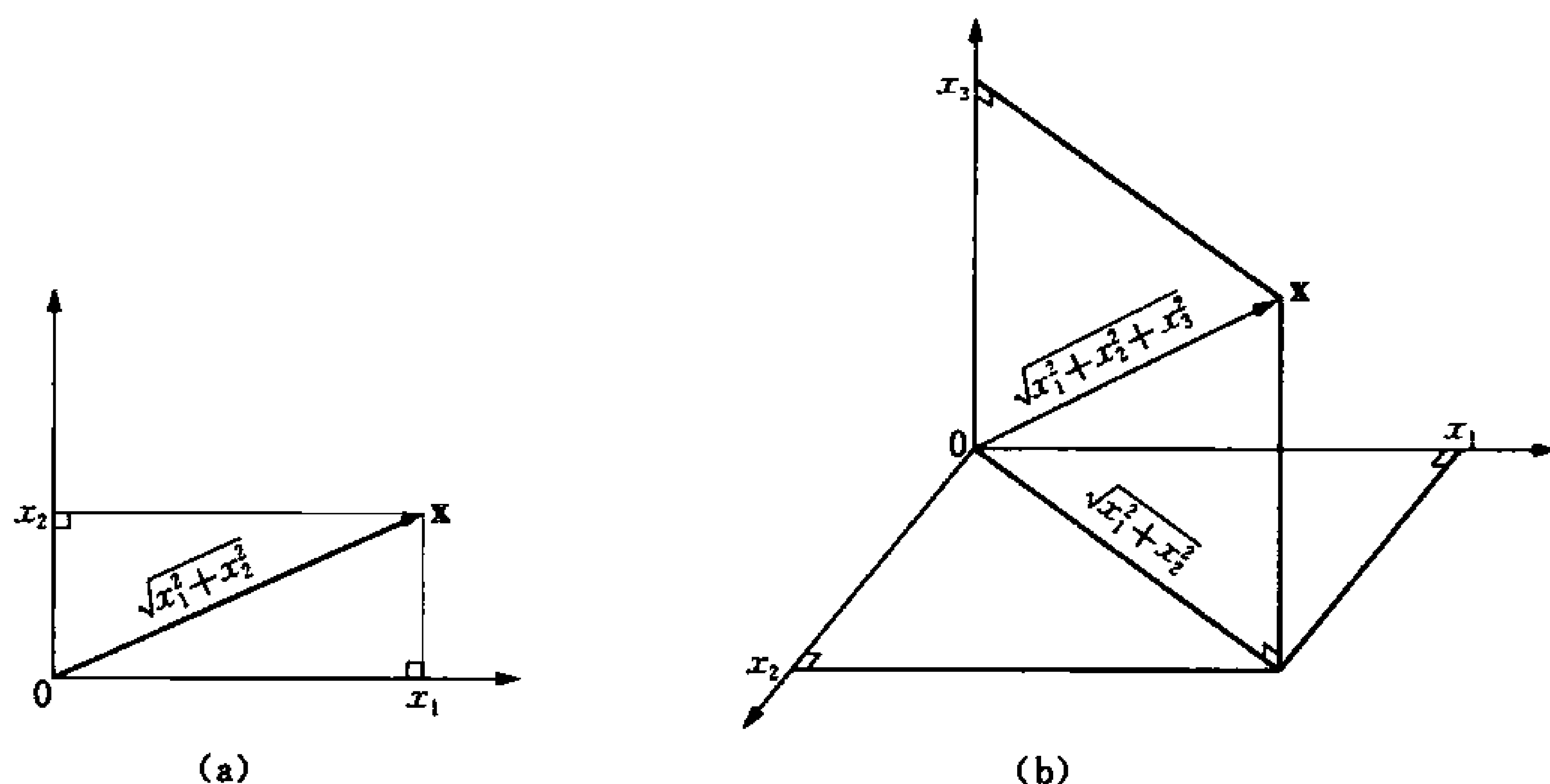
图 1.3 向量 $\mathbf{x}_1-\mathbf{x}_2$ 与向量 $\mathbf{x}_2-\mathbf{x}_1$

在图 1.3 中，向量 $\mathbf{x}_1-\mathbf{x}_2$ 的差表示为以 $\mathbf{x}_2$ 的终点为起点、$\mathbf{x}_1$ 的终点为终点的向量，那么，如果是求向量 $\mathbf{x}_2-\mathbf{x}_1$ 的差，则该向量的方向为从 $\mathbf{x}_1$ 到 $\mathbf{x}_2$。

向量 $\mathbf{x}$ 的长度用 $\|\mathbf{x}\|$ 来表示，等于它坐标平方和的算数平方根：

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

该方程在二维平面中遵循勾股定理，如图 1.4(a)所示。该结果还可以延伸至三维空间坐标中，如图 1.4(b)所示。向量 $\mathbf{x}_1$ 和向量 $\mathbf{x}_2$ 的距离为两个向量终点的距离，表示为 $\|\mathbf{x}_1-\mathbf{x}_2\| = \|\mathbf{x}_2-\mathbf{x}_1\|$（如图 1.3）。



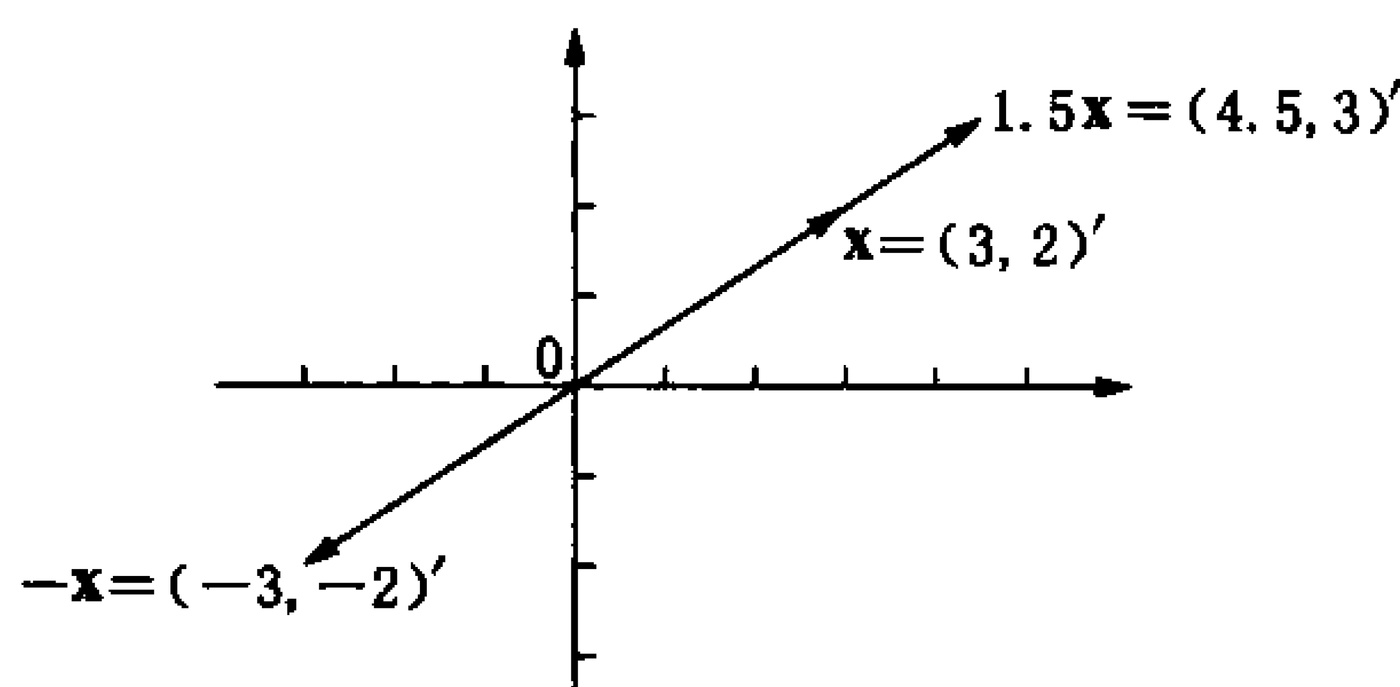
注：其中，(a)与(b)分别为向量长度在二维和三维空间的表示。

图 1.4　向量长度是其坐标平方和的平方根(表示为 $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$)

纯量 a 和向量 $\mathbf{x}$ 的乘积向量 $a\mathbf{x}$ 长度为 $|a| \times \|\mathbf{x}\|$，证明过程如下：

$$\begin{aligned}\|a\mathbf{x}\| &= \sqrt{\sum (ax_i)^2}\\ &= \sqrt{a^2 \sum x_i{}^2}\\ &= |a| \times \|\mathbf{x}\|\end{aligned}$$

如果纯量 a 为正，那么向量 a**x** 与向量 **x** 同向；如果 a 为负，那么向量 $a\mathbf{x}$ 与向量 **x** 共线但是方向相反。向量 $-\mathbf{x}$ 可以看做纯量(−1)和向量 **x** 的乘积，因此，向量 $-\mathbf{x}$ 的长度与 **x** 相同，只是方向相反。这些结果我们都可以在图 1.5 中看到。



注：其中，向量 $a\mathbf{x}$ 与向量 **x** 共线。如果 $a > 0$，那么向量 $a\mathbf{x}$ 与向量 **x** 同向；如果 $a < 0$，那么向量 $a\mathbf{x}$ 与向量 **x** 反向。

图 1.5　向量 ax 在二维坐标平面内的表示

第 3 节 | 向量空间与子空间

n 维向量空间是所有向量 $\mathbf{x} = (x_1, x_2, \cdots, x_n)'$ 的无限组合，且其坐标 x_i 可以是任意实数，因此我们可知，一维向量空间即一条直线，二维向量空间为一个平面，等等。

n 维向量空间的子空间是由空间中含有 k 个向量 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ 的向量空间子集 $\mathbf{y}$ 生成的，该生成集合 $\mathbf{y}$ 的线性组合形式为：

$$\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$$

向量集 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ 分布于整个子空间，我们知道，其实每个 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$ 都是一个由 n 个坐标组成的向量，也就是，$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ 是 k 个向量的集合，而不是一个包含 k 个坐标的向量。

如果该向量集合 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ 中的任何一个向量都无法表示为其他任意向量的线性组合，那么，我们说该向量是几何上线性独立的。

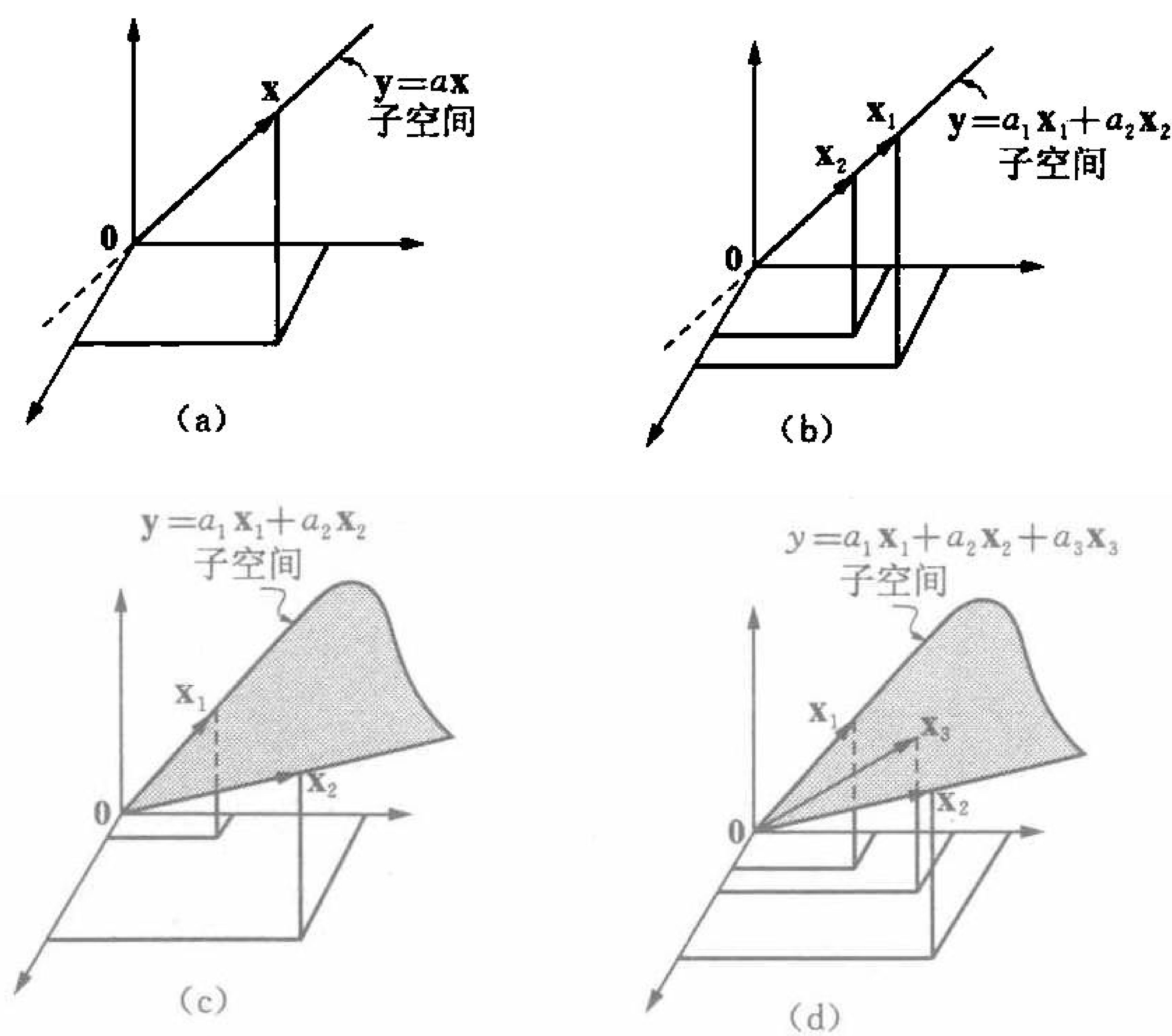
$$\mathbf{x}_j = a_1\mathbf{x}_1 + \cdots + a_{j-1}\mathbf{x}_{j-1} + a_j\mathbf{x}_j + \cdots + a_k\mathbf{x}_k \qquad [1.6]$$

其中，一些常数 a_l 可为 0。同样，我们可以说，如果不存在不全为 0 的常数 $b_1, b_2, \cdots, b_k$ 使得

$$b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k = \underset{(n\times 1)}{\mathbf{0}} \qquad [1.7]$$

那么，该向量集合线性独立。方程 1.6 和方程 1.7 则被称为"线性相关"或者"共线性方程"。当向量集合符合这两个方程时，则我们称该集合为"线性相关集合"。注意，由于 $\mathbf{0} = 0\mathbf{x}$，因此零向量与任何向量都存在线性相关的关系。

子空间的维度是由最大的线性独立子集内的向量个数决定的。因此，由向量集合 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ 生成的子空间维度不会超过 k 和 n。这些在向量空间中的关系可在图 1.6 的三维坐标系中体现出来。图 1.6(a)表示由一个非零向量 $\mathbf{x}$ 生成的一维子空间(直线)；图 1.6(b)表示由经 $\mathbf{x}_1$、$\mathbf{x}_2$ 两个共线向量组成的



注：(a)由一个非零向量生成的一维子空间(一条直线)；(b)由两个共线向量生成的一维子空间；(c)由两个线性独立的向量生成的二维子空间(一个平面)；(d)由三条线性相关但是两两之间线性独立的向量生成的二维子空间。其中，(c)和(d)中生成的平面可以无限延伸，将平面画在 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 之间是表达的需要。

图 1.6 三维空间的向量集生成的子空间

一维子空间；图 1.6(c)表示由两个线性独立的向量 $\mathbf{x}_1$、$\mathbf{x}_2$ 组成的二维子空间(平面)；最后，图 1.6(d)表示由三个线性相关的向量 $\mathbf{x}_1$、$\mathbf{x}_2$、$\mathbf{x}_3$ 生成的二维子空间。在最后一个例子中，任意一个向量都会落在由其他两个向量组成的平面中。

一个线性独立的向量集 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$，如图 1.6(a)中的 $\{\mathbf{x}\}$ 和图 1.6(c)中的 $\{\mathbf{x}_1, \mathbf{x}_2\}$，均可以看做向量集所扩张出的子空间的基。空间内的每个向量都能以唯一的方式表达成这些基向量的线性组合：

$$\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k$$

常数 $c_1, c_2, \cdots, c_k$ 被称为“$\mathbf{y}$ 的坐标值”。因为 $\mathbf{0} = 0\mathbf{x}_1 + 0\mathbf{x}_2 + \cdots + 0\mathbf{x}_k$，所以零向量可以存在于任何子空间。

一个二维子空间的向量坐标可以根据向量加法中的平行四边形法则找出(如图 1.7)。我们还可以通过线性联立方程组得到具体坐标值，其中，$c_1, c_2, \cdots, c_k$ 为未知量。

$$\underset{(n\times 1)}{\mathbf{y}} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k$$

$$= [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k]\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix}$$

$$= \underset{(n\times k)}{\mathbf{X}}\ \underset{(k\times 1)}{\mathbf{c}}$$

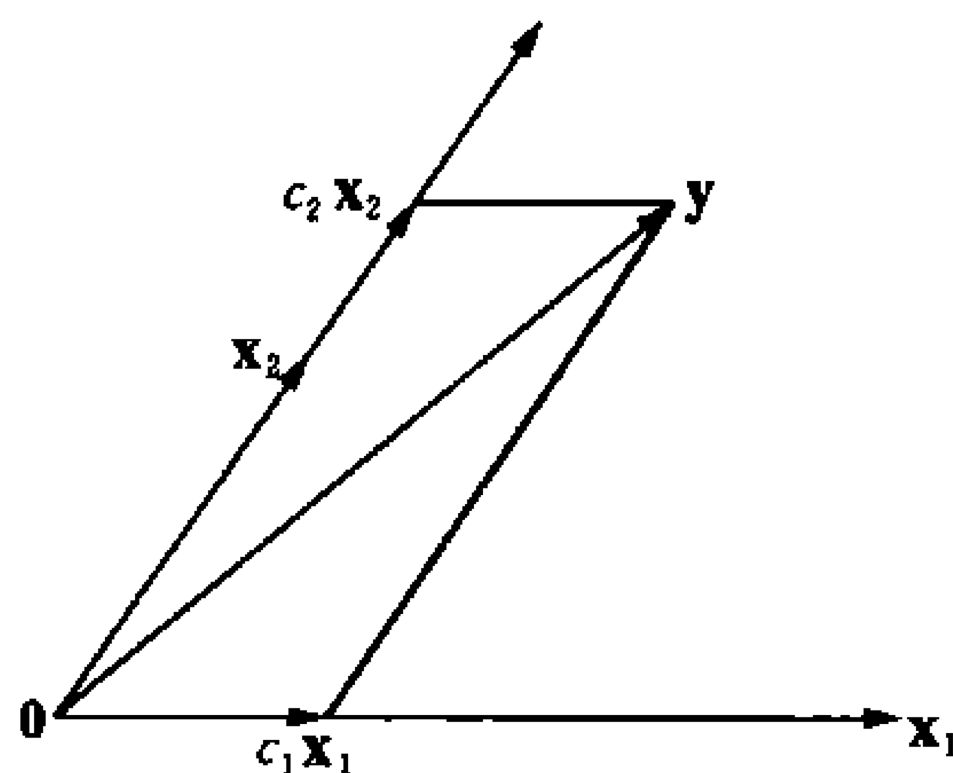


图 1.7 通过向量加法的平行四边形法则得到的以 $\{\mathbf{x}_1, \mathbf{x}_2\}$ 为基的向量 $\mathbf{y}$ 的坐标

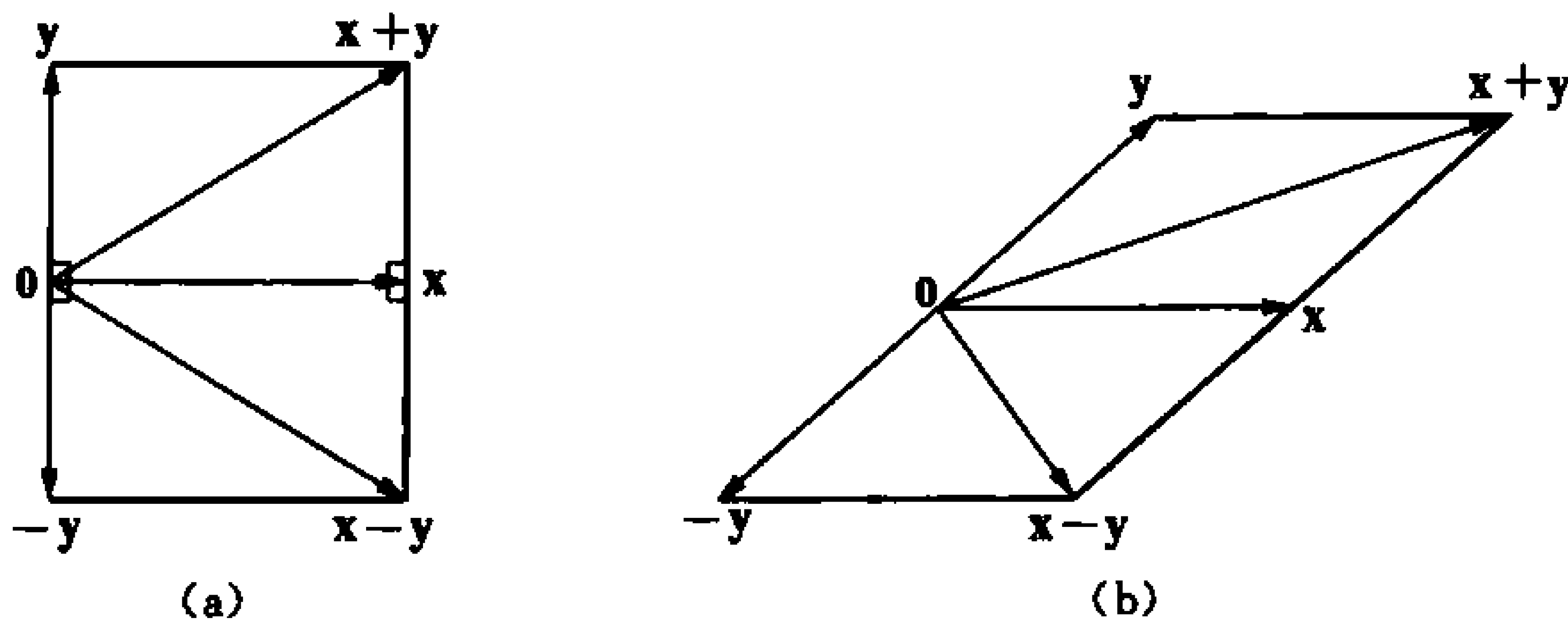
当向量集 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ 中的向量是线性独立的时候，矩阵 **X** 为列满秩矩阵，此时，方程组有唯一解。有关轶的概念和系统线性联立方程组的解法，我们会在之后介绍。

正交与正交投影

我们知道，两个向量的内积等于它们对应坐标的乘积之和：

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

如果两个向量 **x**、**y** 正交(垂直)，那么它们的内积为 0。有关正交的基本几何向量可参见图 1.8。尽管向量 **x** 和 **y** 均存在于一个 n 维空间中(因此有些我们可能无法直接观测到)，但按照惯例，我一般将其画在一个二维平面坐标里。[3] 如图 1.8(a)所示，当向量 **x** 和 **y** 正交时，顶点分别为(**0**, **x**, **x**+**y**)和



注：(a)当向量 **x**, **y** 正交时，它们的内积 **x**·**y** 等于 **0**；(b)当两向量不正交，那么它们的内积不等于 **0**。

图 1.8 正交的基本几何向量

(**0**, **x**, **x**−**y**)的两个直角三角形是全等的。因此，$\|\mathbf{x}+\mathbf{y}\| = \|\mathbf{x}-\mathbf{y}\|$。由于向量的长度为该向量与其本身的内积的平方根，于是我们有：

$$(\mathbf{x}+\mathbf{y})\cdot(\mathbf{x}+\mathbf{y})=(\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})$$

$$\mathbf{x}\cdot\mathbf{x}+2\mathbf{x}\cdot\mathbf{y}+\mathbf{y}\cdot\mathbf{y}=\mathbf{x}\cdot\mathbf{x}-2\mathbf{x}\cdot\mathbf{y}+\mathbf{y}\cdot\mathbf{y}$$

$$4\mathbf{x}\cdot\mathbf{y}=0$$

$$\mathbf{x}\cdot\mathbf{y}=0$$

相反,当 $\mathbf{x}$ 和 $\mathbf{y}$ 不正交时,那么 $\|\mathbf{x}+\mathbf{y}\| \neq \|\mathbf{x}-\mathbf{y}\|$,则 $\mathbf{x}\cdot\mathbf{y}\neq 0$。

向量 $\mathbf{y}$ 在向量 $\mathbf{x}$ 上的正交投影可看做向量 $\mathbf{x}$ 与一个纯量的乘积,那么,$(\mathbf{y}-\hat{\mathbf{y}})$ 与 $\mathbf{x}$ 正交。正交投影的几何表示请见图 1.9。通过平行四边形法则(见图 1.10),$\hat{\mathbf{y}}$ 的终点是向量 $\mathbf{x}$ 方向上与向量 $\mathbf{y}$ 的终点距离最近的点。为了找到正确的纯量 b,我们有:

$$\mathbf{x}\cdot(\mathbf{y}-\hat{\mathbf{y}})=\mathbf{x}\cdot(\mathbf{y}-b\mathbf{x})=0$$

因此,$\mathbf{x}\cdot\mathbf{y}-b\mathbf{x}\cdot\mathbf{x}=0$,那么,$b=(\mathbf{x}\cdot\mathbf{y})/(\mathbf{x}\cdot\mathbf{x})$。

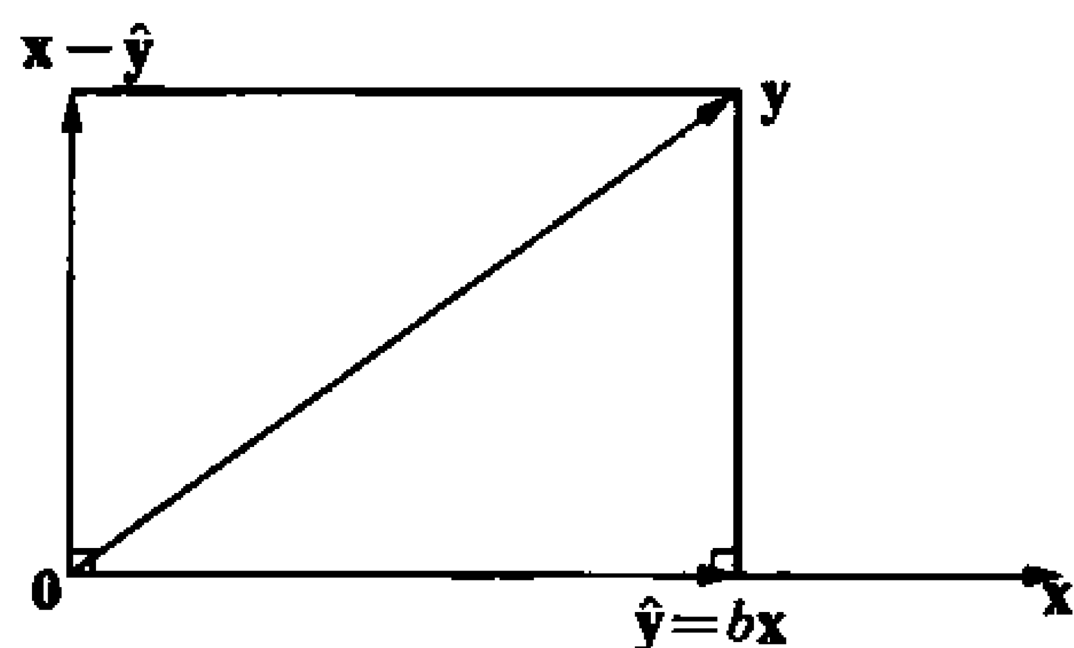


图 1.9 向量 $\mathbf{y}$ 在向量 $\mathbf{x}$ 上的正交投影 $\hat{\mathbf{y}}=b\mathbf{x}$

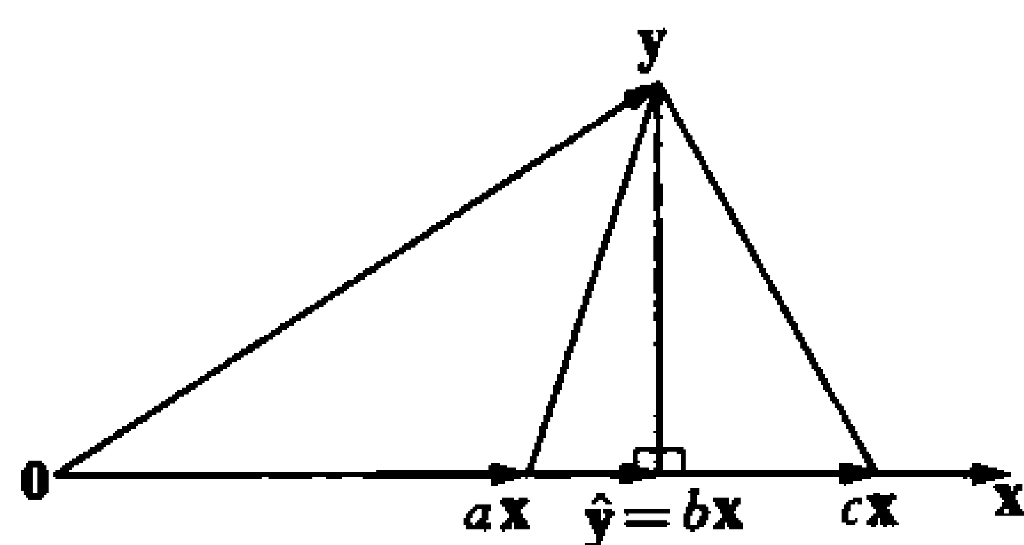


图 1.10 向量 $\mathbf{y}$ 在向量 $\mathbf{x}$ 上的正交投影 $\hat{\mathbf{y}}=b\mathbf{x}$(其终点为与向量 $\mathbf{y}$ 的终点距离最近的点)

当正交的定义延伸到矩阵中时,则有:若矩阵 $\mathbf{X}$ 的列向量两两正交,即当 $\mathbf{X}'\mathbf{X}$ 为对角矩阵时[4],矩阵 $\underset{(n\times k)}{\mathbf{X}}$ 为正交矩阵。所以,如果矩阵 $\mathbf{X}$ 为正交矩阵,其符合 $\mathbf{X}'\mathbf{X}=\mathbf{I}$。

通过向量 $\mathbf{y}$ 在 $\mathbf{x}$ 上的正交投影可以得到两向量夹角的余弦值进而算出两个向量之间的夹角。由于余弦函数在 $w=0$

上中心对称，因此在任意方向上测量夹角都可以。这里，我简单地把所有夹角都视为正值。有关余弦及其他三角函数的讨论请见后文。我将夹角类型大概分为两种：两向量之间的夹角在0°和90°之间；两向量夹角在90°和180°之间。[5]对于第一种类型：

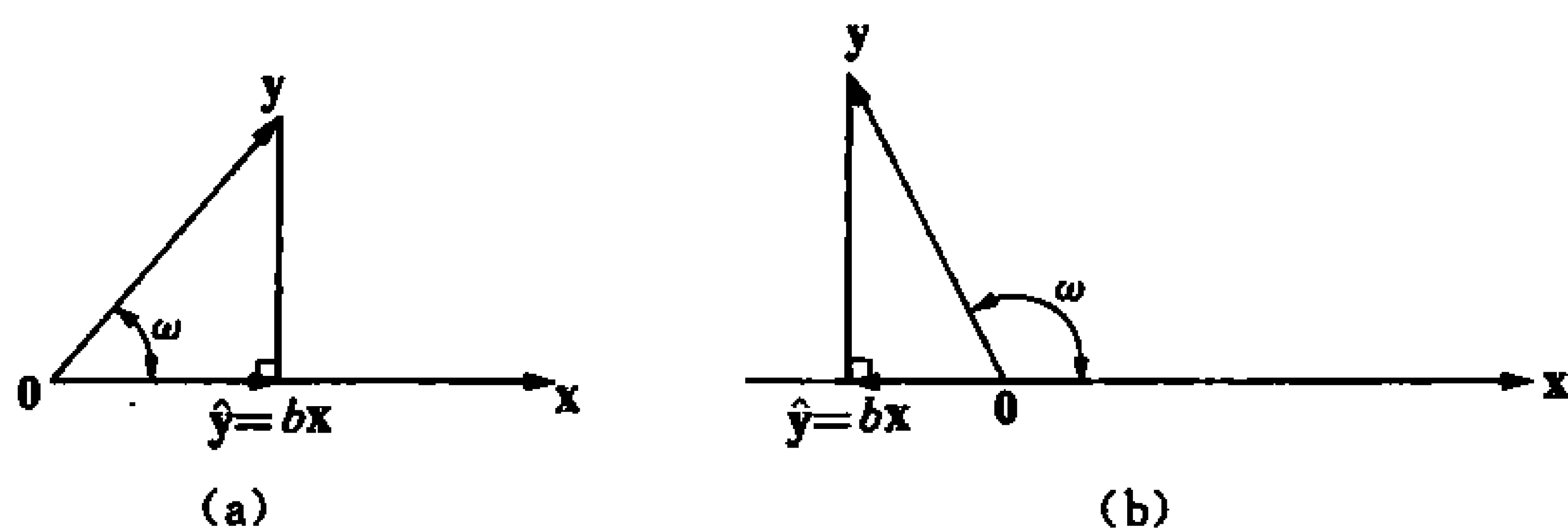
$$\cos w = \frac{\| \hat{\mathbf{y}} \|}{\| \mathbf{y} \|} = \frac{b \| \mathbf{x} \|}{\| \mathbf{y} \|} = \frac{\mathbf{x} \cdot \mathbf{y}}{\| \mathbf{x} \|^2} \times \frac{\| \mathbf{x} \|}{\| \mathbf{y} \|}$$

$$= \frac{\mathbf{x} \cdot \mathbf{y}}{\| \mathbf{x} \| \times \| \mathbf{y} \|}$$

对于第二种类型：

$$\cos w = -\frac{\| \hat{\mathbf{y}} \|}{\| \mathbf{y} \|} = \frac{b \| \mathbf{x} \|}{\| \mathbf{y} \|} = \frac{\mathbf{x} \cdot \mathbf{y}}{\| \mathbf{x} \| \times \| \mathbf{y} \|}$$

对于以上两种情况，向量 $\mathbf{y}$ 在 $\mathbf{x}$ 上的正交投影的 b 的符号反映了 $\cos w$ 的符号。



注：(a) $0° < w < 90°$；(b) $90° < w < 180°$。

图1.11　向量x与y的夹角

在一个由向量集合 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ 扩张出来的子空间中，向量 $\mathbf{y}$ 的正交投影可表示为向量 $\mathbf{x}_j (j = 1, 2, \cdots, k)$ 的线性组合。因此，$(\mathbf{y} - \hat{\mathbf{y}})$ 与该向量集里的所有向量 $\mathbf{x}_j$ 都正交。

$$\hat{\mathbf{y}} = b_1 \mathbf{x}_1 + b_2 \mathbf{x}_2 + \cdots + b_k \mathbf{x}_k$$

如果 $k=2$，该正交投影的几何表述可参见图 1.12。在由该向量集扩张出来的子空间里，向量 $\hat{\mathbf{y}}$ 的终点为在向量 $\mathbf{x}$ 方向上与向量 $\mathbf{y}$ 的终点距离最近的点。

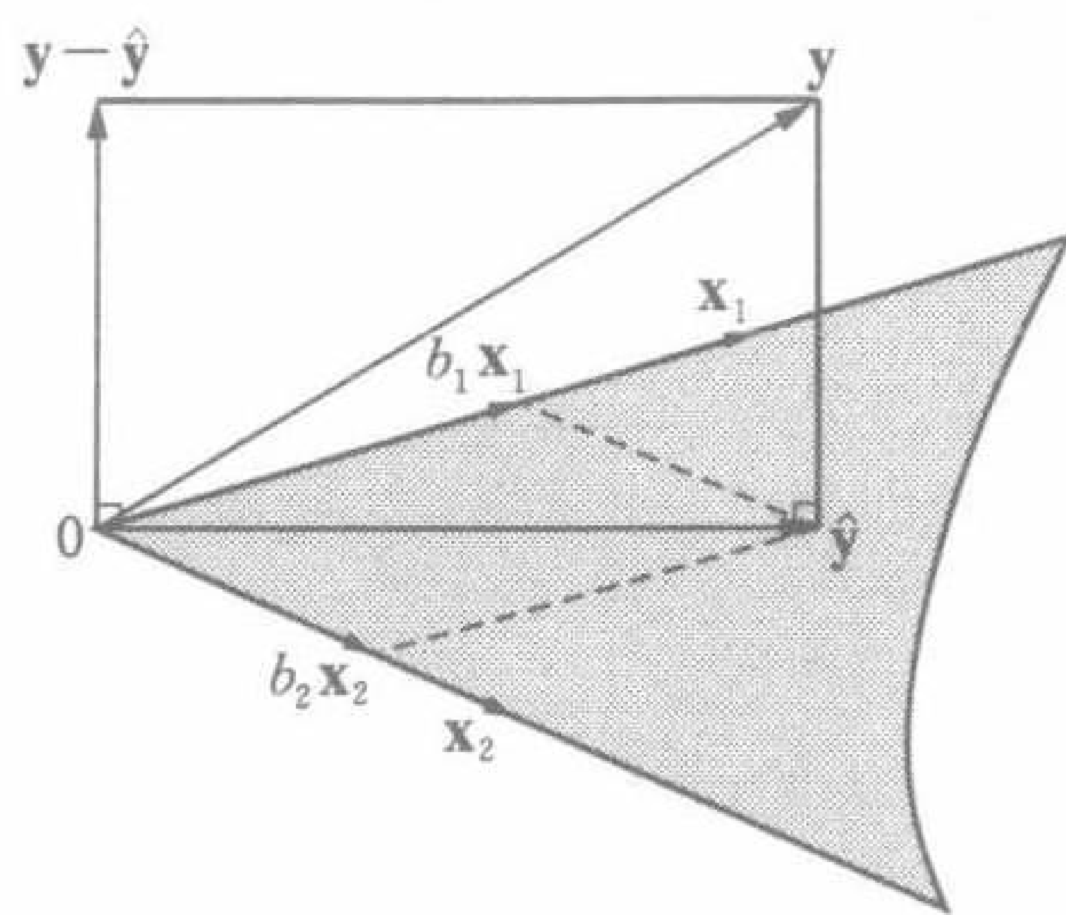


图 1.12　向量 y 在由向量 $\mathbf{x}_1$、$\mathbf{x}_2$ 扩张出来的子空间(平面)上的正交投影 $\hat{\mathbf{y}}$

我们用一个向量 $\mathbf{b}$ 包含所有常数 b_j，同时把所有向量 $\mathbf{x}_j$ 放入一个 $(n\times k)$ 矩阵 $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_k]$ 中，因此，我们有 $\hat{\mathbf{y}}=\mathbf{X}\mathbf{b}$。根据正交投影定义，得到：

$$\mathbf{x}_j\cdot(\mathbf{y}-\hat{\mathbf{y}})=\mathbf{x}_j\cdot(\mathbf{y}-\mathbf{X}\mathbf{b})=0\ (j=1,\cdots,k)\qquad[1.8]$$

同理，$\mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{b})=\mathbf{0}$，$\mathbf{X}'\mathbf{y}=\mathbf{X}'\mathbf{X}\mathbf{b}$。只要 $\mathbf{X}'\mathbf{X}$ 为非奇异矩阵，我们就可以找到符合该方程的唯一的 $\mathbf{b}$。对于基向量，只要 $\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_k\}$ 线性独立，则 $\mathbf{X}'\mathbf{X}$ 为非奇异矩阵，$\mathbf{b}$ 有唯一解，否则，$\mathbf{b}$ 的解不唯一。

有关正交投影在最小二乘线性回归中的应用非常直接。假设图 1.9 和图 1.11 中的向量 $\mathbf{x}$ 是一个简单回归里的自变量，向量 $\mathbf{y}$ 为因变量，对于 $\mathbf{x}$ 和 $\mathbf{y}$ 我们都用(每个变量与其均值的)偏差来表示，则有 $\mathbf{x}=\{X_i-\bar{X}\}$，$\mathbf{y}=\{Y_i-\bar{Y}\}$。那么，$\hat{\mathbf{y}}=b\mathbf{x}$ 即 Y 对 X 进行最小二乘线性回归后，通过 Y 值拟合得到的平均偏差向量；b 为斜率，$\mathbf{y}-\hat{\mathbf{y}}$ 为最小二乘残差向量。根据平行四边形法则，我们发现，Y 的总平方和可以分

解为回归平方和和残差平方和，即：

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y}-\hat{\mathbf{y}}\|^2$$

或者叫做回归的“方差分析”。那么，$\mathbf{x}$ 和 $\mathbf{y}$ 之间的相关系数 r 就是它们平均偏差向量夹角的余弦值。

同样，在一个多元回归中，我们假设 $\mathbf{y}$ 为因变量的平均偏差向量，$\mathbf{x}_1$ 和 $\mathbf{x}_2$ 为两个自变量的平均偏差向量，那么，Y 对 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的最小二乘线性回归则如图1.12 所示。其中，b_1 和 b_2 为两个自变量的偏回归系数。由原点、$\mathbf{y}$ 及 $\hat{\mathbf{y}}$ 组成的直角三角形给出了多元回归中的方差分析。$\mathbf{y}$ 与 $\hat{\mathbf{y}}$ 之间夹角的余弦值则为回归得出的 R，即观测的 Y 与回归拟合出的 Y 的相关性大小。

第 4 节 | 矩阵的秩及线性联立方程组的解法

矩阵的秩

（$m \times n$）矩阵 $\mathbf{A}$ 的行空间是 n 维向量空间的子空间，该子空间是由矩阵 $\mathbf{A}$ 的 m 行向量生成的。矩阵 $\mathbf{A}$ 的秩即行空间维数，换句话说，矩阵 $\mathbf{A}$ 的秩是最大线性独立行数值。它遵循 $\text{rank}(\mathbf{A}) \leqslant \min(m, n)$。

例如，矩阵的行空间

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

包含的所有向量为：

$$\begin{aligned} \mathbf{x}' &= a[1, 0, 0] + b[0, 1, 0] \\ &= [a, b, 0] \end{aligned}$$

该子空间维数为 2，因此，$\text{rank}(\mathbf{A}) = 2$。

如果一个矩阵为行简化阶梯形矩阵（RREF），那么，它必须符合以下标准：

R1：如果矩阵中包含零行，零行必须排在非零行后面。

R2：从左到右，每个非零行的首非零元都为1。

R3：若第 m 行的首非零元位置在 k 列，那么，第 $m+1$ 行的首非零元位置则在 $k+1$ 列。

R4：首非零元所在列的其他元均为0。

方程1.9形象地列出了行简化阶梯形矩阵，其中，"$*$"号表示所在元素的任意值：

$$\begin{bmatrix} 0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * & 0 & * & \cdots & * \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & & & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & * & \cdots & * \\ \hline 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{matrix} \\ \text{非零行} \\ \\ \\ \text{零行} \\ \\ \\ \end{matrix} \qquad [1.9]$$

行简化阶梯形矩阵(RREF)的秩和矩阵中的非零行数目相等，首非零元所在列的其他元均为0的性质，保证了任意非零行不可能成为其他非零行的线性组合。

通过一系列初等行变换，我们可以把一个矩阵变为RREF。例如，

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

1. 第一行除以-2，

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

2. 第二行减去第一行乘以4，

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & -1 & 4 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

3. 第三行减去第一行乘以6，

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & -1 & 4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

4. 第二行乘以-1，

$$\begin{bmatrix} 1 & 0 & \frac{1}{2} & -1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

5. 第一行减去第二行乘以$\frac{1}{2}$，

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & -2 & 8 \end{bmatrix}$$

6. 第三行加上第二行乘以 2，

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

由于矩阵 **A** 中有一个零行，我们知道，零行可以写成其他行的线性组合，所以矩阵 **A** 的秩等于其行数减 1，其值等于矩阵 **A** 的行简化阶梯形矩阵(RREF)——$\mathbf{A}_R$ 的秩。因此，我们可知，初等行变换不会改变一个矩阵的秩。

一个非奇异方形矩阵的 RREF 是一个单位矩阵，因此，非奇异方形矩阵的秩又等于其阶数。相反，一个奇异矩阵的值比其阶数小。

之前我们定义矩阵 **A** 的秩为其行空间的维度。其实，矩阵 **A** 的秩与其列空间的维数也相等，换句话说，矩阵 **A** 的秩又等于矩阵 **A** 中线性独立的列数。

线性联立方程组

含有 n 个未知数的 m 个线性方程组用矩阵形式可表达为：

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{b}} \qquad [1.10]$$

其中，矩阵 **A** 是由未知数的系数组成的，向量 **b** 是由方程等号右边的常数项组成的，**x** 为未知数向量。假设方程的数目和未知数的数目相等，即 $m = n$，或者矩阵 **A** 为非奇异矩阵，那么，方程 1.10 有唯一解，即 $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$。

同理，如果 **A** 为奇异矩阵，那么 **A** 就可以通过一系列初等行变换被转化为 RREF：

$$\mathbf{A}_R = \mathbf{E}_P \cdots \mathbf{E}_2 \mathbf{E}_1 \mathbf{A} = \mathbf{E}\mathbf{A}$$

通过对方程左边和右边同时应用行操作，则得到：

$$\mathbf{E}\mathbf{A}\mathbf{x} = \mathbf{E}\mathbf{b}$$

$$\mathbf{A}_R \mathbf{x} = \mathbf{b}_R \qquad [1.11]$$

其中，$\mathbf{b}_R \equiv \mathbf{E}\mathbf{b}$。因此方程 1.10 和方程 1.11 是等价的。

以 r 表示矩阵 $\mathbf{A}$ 的秩。$r < n$（考虑如果矩阵 $\mathbf{A}$ 为奇异矩阵），$\mathbf{A}_R$ 包含 r 个非零行和 $n-r$ 个零行。如果矩阵 $\mathbf{A}_R$ 的任意零行在向量 $\mathbf{b}_R$ 中的对应元不为 0，那么，该方程组是不一致的，我们称这样的方程组为“超定方程组”，因为该方程组中存在自相矛盾的方程。

$$0x_1 + 0x_2 + \cdots + 0x_n = b \neq 0$$

如果矩阵 $\mathbf{A}_R$ 的任意零行在向量 $\mathbf{b}_R$ 中的对应元为 0，则该方程组是一致的，但此时该方程组却有无穷多个解，其中 $n-r$ 个未知数可以取任意值，这 $n-r$ 个未知数又决定了其他 r 个未知数也会有无穷多个解。我们称这样的方程组为“欠定方程组”。假设方程的数目小于未知数的数目，即 $m < n$，那么，r 必然小于 n，该方程组既可能是超定方程组（如果矩阵 $\mathbf{A}_R$ 的零行所对应的向量 $\mathbf{b}_R$ 的元非零），也可能是欠定方程组（如果方程组是一致的）。我们可以考虑以下方程组：

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

将等号右边的常数矩阵加入该矩阵后得到：

$$\left[\begin{array}{cccc|c} -2 & 0 & -1 & 2 & 1 \\ 4 & 0 & 1 & 0 & 2 \\ 6 & 0 & 1 & 2 & 5 \end{array}\right]$$

将左边系数矩阵变为行简化矩阵的步骤为：

1. 第一行除以-2，

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 4 & 0 & 1 & 0 & 2 \\ 6 & 0 & 1 & 2 & 5 \end{array}\right]$$

2. 第二行减去 4 乘以第一行，第三行减去 6 乘以第一行，

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 0 & 0 & -1 & 4 & 4 \\ 0 & 0 & -2 & 8 & 8 \end{array}\right]$$

3. 第二行乘以-1，

$$\left[\begin{array}{cccc|c} 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} \\ 0 & 0 & 1 & -4 & -4 \\ 0 & 0 & -2 & 8 & 8 \end{array}\right]$$

4. 第一行减去$\frac{1}{2}$乘以第二行，第三行加上 2 乘以第二行，

$$\left[\begin{array}{cccc|c} 1\swarrow & 0 & 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1\swarrow & -4 & -4 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$ （首非零元用箭头标出）

写成方程组形式，我们得到：

$$x_1 + x_4 = \frac{3}{2}$$

$$x_3 - 4x_4 = -4$$

$$0x_1 + 0x_2 + 0x_3 + 0x_4 = 0$$

第三个方程没有提供任何有用的信息，但是它说明了"原方程组是一致的"。前两个方程暗示了未知数 x_1 和 x_4 可以取任意值（我们用 x_2^* 和 x_4^* 表示），那么，x_1 和 x_3 可表示为：

$$x_1 = \frac{3}{2} - x_4^*$$

$$x_3 = -4 + 4x_4^*$$

因此，任意向量

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} - x_4^* \\ x_2^* \\ -4 + 4x_4^* \\ x_4^* \end{bmatrix}$$

为原方程组的解。

现在，我们考虑另一个方程组：

$$\begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

将向量 b 合并到系数矩阵 A 中进行初等行变换后，我们得到的行简化矩阵为：

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & -4 & -2 \\ 0 & 0 & 0 & 0 & 2 \end{array}\right]$$

最后一个方程 $0x_1 + 0x_2 + 0x_3 + 0x_4 = 2$ 是自相矛盾的，因此，原方程组无解。

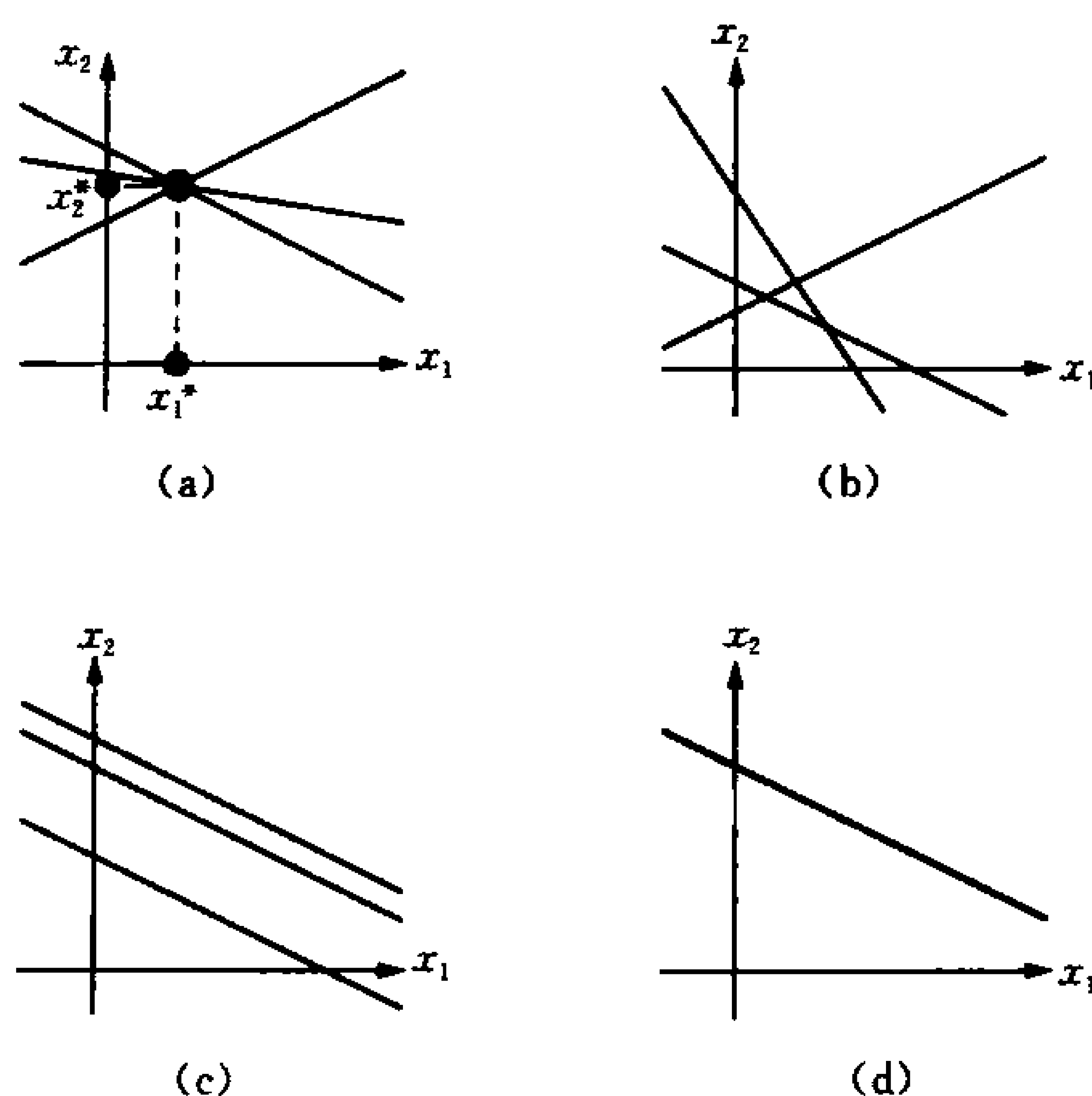
假设方程组中方程的数目大于未知数的数目，即 $m > r$，如果矩阵 A 为列满秩矩阵（$r = n$），那么矩阵 $\mathbf{A}_R$ 包含 n 阶单位矩阵和 $m - r$ 个零行。若方程组是一致的，那么，该方程组有唯一解，否则，该方程组为超定方程组。当 $r < n$ 时，方程组既可能为超定方程组，也可能为欠定方程组。

我们可以在一个二元方程组[6]里证明以上论述：

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$
$$a_{31}x_1 + a_{32}x_2 = b_3$$

每个方程都可以在一个二维平面坐标系里表示，其中，坐标轴由两个未知数构成（如图 1.13）。如果三条直线相交于一点，如图 1.13(a)，那么方程组有唯一解——两个未知数（x_1^*、x_2^*）同时满足三个方程。如果三条直线没有相交于一点，如图 1.13(b)和图 1.13(c)，那么两个未知数无法同时满足三个方程，因此该方程组为超定方程组。最后，如果三条直线重合，如图 1.13(d)，无论未知数

取什么值，都可以满足三个方程，此时，方程组被称为“欠定方程组”。



注：(a) 唯一解；(b)和(c)超定方程组；(d) 欠定方程组(三条直线重合)。

图 1.13 含有两个未知数的三个线性方程

如果等号右边的向量 **b** 在线性联立方程组里为零向量时，方程组被称为“齐次方程组”：

$$\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}} = \mathbf{0} \qquad [1.12]$$

那么，平凡解 $\mathbf{x} = \mathbf{0}$ 总是符合齐次方程组，因此，方程组不可能不一致。通过上文的介绍，我们知道，非平凡解只有当 $\text{rank}(\mathbf{A}) < n$ 时，即方程组为欠定方程组时才存在。

表 1.1 总结了有关线性联立方程不同情况下的解。[7]

线性联立方程在统计上被广泛运用，例如，我们熟悉的最小二乘回归分析。

表 1.1 含有 n 个未知数和 m 个线性联立方程的解

方程个数	$m<n$	$m=n$		$m>n$	
系数矩阵的秩	$r<n$	$r<n$	$r=n$	$r<n$	$r=n$
一般方程系统					
一致	欠定	欠定	唯一解	欠定	唯一解
不一致	超定	超定	—	超定	超定
齐次方程系统					
一致	非平凡解	非平凡解	平凡解	非平凡解	平凡解

广义逆矩阵

我们了解到只有方形非奇异矩阵才有逆阵。那么，对于所有矩阵，包括奇异矩阵及长方形矩阵，它们拥有的是广义逆矩阵，广义逆矩阵在统计学中非常有用，例如，在介绍线性统计模型时非常有用。[8]

$(m\times n)$ 阶矩阵 $\mathbf{A}$ 的广义逆矩阵为 $(n\times m)$ 矩阵 $\mathbf{A}^-$，其满足方程：

$$\mathbf{A}\mathbf{A}^-\mathbf{A}=\mathbf{A}$$

请注意，$\mathbf{A}^-$ 是一个广义逆矩阵，而不是矩阵 $\mathbf{A}$ 的广义逆矩阵。除非 $\mathbf{A}$ 是方形非奇异矩阵（在这种情况下，$\mathbf{A}^-=\mathbf{A}^{-1}$），否则广义逆矩阵就不是唯一的。[9]

许多方法可以帮助我们找到矩阵的广义逆矩阵，例如，高斯消去法。我们先通过初等行变换把矩阵 $\mathbf{A}$ 变为 RREF：

$$\mathbf{E}\mathbf{A}=\mathbf{E}_P\cdots\mathbf{E}_2\mathbf{E}_1\mathbf{A}=\mathbf{A}_R \qquad [1.13]$$

其中，$\mathbf{E}=\mathbf{E}_P\cdots\mathbf{E}_2\mathbf{E}_1$ 是一个 $(m\times m)$ 的非奇异矩阵。再通过

第二类、第三类初等列变换(转置是不必要的,因为 $\mathbf{A}_R$ 中所有的首非零元已经为 1),我们进一步将 $\mathbf{A}_R$ 简化为标准形式:

$$\underset{(m\times n)}{\mathbf{A}_C} \equiv \mathbf{A}_R\mathbf{E}^* = \mathbf{A}_R\mathbf{E}_1^*\mathbf{E}_2^*\cdots\mathbf{E}_q^* = \begin{bmatrix} \mathbf{I}_r & \underset{(r\times n-r)}{\mathbf{0}} \\ \underset{(m-r\times r)}{\mathbf{0}} & \underset{(m-r\times n-r)}{\mathbf{0}} \end{bmatrix} \tag{1.14}$$

其中,$\mathbf{E}^* = \mathbf{E}_1^*\mathbf{E}_2^*\cdots\mathbf{E}_q^*$ 是一个 $(n\times n)$ 的非奇异矩阵,左上角单位矩阵的阶数 r 是矩阵 $\mathbf{A}$ 的秩,其他所有零矩阵可有可无。因为,如果 $\mathbf{A}$ 是一个 n 阶非奇异矩阵,则 $r = n$,那么在这里,我们就不需要零矩阵。

将方程 1.13 和方程 1.14 合并,得到:

$$\mathbf{A}_C = \mathbf{E}\mathbf{A}\mathbf{E}^* \tag{1.15}$$

那么,$\mathbf{A}$ 的广义逆矩阵为[10]:

$$\mathbf{A}^- = \mathbf{E}^*\mathbf{A}_C'\mathbf{E}$$

现在我们考虑矩阵:

$$\mathbf{A} = \begin{bmatrix} -2 & 0 & -1 & 2 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 1 & 2 \end{bmatrix}$$

在上文中,我们将该矩阵变为行简化矩阵后,得到:

$$\mathbf{A}_R = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

通过交换第二列、第三列,把第四列元素归零,将矩阵化为标

准形式后得到：

$$\mathbf{A}_C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

将所有的初等行列变换写成矩阵，我们得到：

$$\mathbf{E} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -2 & -1 & 0 \\ -1 & -2 & 1 \end{bmatrix}$$

$$\mathbf{E}^* = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

通过以上矩阵，

$$\mathbf{A}^- = \mathbf{E}^* \mathbf{A}'_C \mathbf{E}$$

$$= \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -2 & -1 & 0 \\ -1 & -2 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ -2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

我们得到的 $\mathbf{A}^-$ 为矩阵 $\mathbf{A}$ 的一个广义逆矩阵。

我们考虑一个含有 n 个未知数和 m 个线行联立方程的方程组，

$$\underset{(m\times n)}{\mathbf{A}}\underset{(n\times 1)}{\mathbf{x}} = \underset{(m\times 1)}{\mathbf{b}}$$

假设该方程组是一致的且为欠定方程组，那么，

$$\mathbf{x}^{*} = \mathbf{A}^{-}\mathbf{b} \qquad [1.16]$$

该方程组有无数解。如果方程组有唯一解，那么，我们可以通过方程 1.16 算出。最后，如果该方程组是超定的，那么，方程 1.16 无法满足原方程组，即方程组无解。因此，我们可以知道，如果方程组是一致的，那么，$\mathbf{AA}^{-}\mathbf{b}=\mathbf{b}$，否则，$\mathbf{AA}^{-}\mathbf{b}\neq\mathbf{b}$。

第 5 节 | 特征值与特征向量

如果 $\mathbf{A}$ 为 n 阶方阵，那么，齐次线性方程组

$$(\mathbf{A}-\lambda\mathbf{I}_n)\mathbf{x}=\mathbf{0} \qquad [1.17]$$

只有在纯量 λ 为某几个特定数值时，它有非平凡解。通过上面的内容，我们知道，当矩阵 $(\mathbf{A}-\lambda\mathbf{I}_n)$ 为奇异矩阵时，方程组存在平凡解，即当满足下列条件时：

$$\det(\mathbf{A}-\lambda\mathbf{I}_n)=0 \qquad [1.18]$$

方程 1.18 称为矩阵 $\mathbf{A}$ 的“特征方程”，λ 为矩阵 $\mathbf{A}$ 的特征值、特征根或者潜伏根。在某一特征值 λ_1 下满足方程 1.17 的向量 $\mathbf{x}_1$ 称为在特征值 λ_1 下，矩阵 $\mathbf{A}$ 的“特征向量”。

为简单起见，我用一个 (2×2) 矩阵的例子来详细解释。对于此例，特征方程可写为：

$$\det\begin{bmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{bmatrix}=0$$

$$(a_{11}-\lambda)(a_{22}-\lambda)-a_{12}a_{21}=0$$

$$\lambda^2-(a_{11}+a_{22})\lambda+a_{11}a_{22}-a_{12}a_{21}=0$$

利用一元二次方程的相关公式来计算其两个平方根，则有[11]：

$$\lambda_1 = \frac{1}{2}\left[a_{11} + a_{22} + \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right]$$

$$\lambda_2 = \frac{1}{2}\left[a_{11} + a_{22} - \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right]$$

[1.19]

如果根号以下部分非负，那么该平方根必为实数。注意，有可能存在 $\lambda_1 + \lambda_2 = a_{11} + a_{22}$（A 的特征值之和等于 A 的迹）和 $\lambda_1\lambda_2 = a_{11}a_{22} - a_{12}a_{21}$（特征值的积等于矩阵 A 的行列式的值）的情况。而且，如果 A 为奇异矩阵，则 λ_2 为 0。

当矩阵 A 为对称矩阵（在特征值和特征向量的统计应用中很常见）时，有 $a_{12} = a_{21}$，方程 1.19 变为：

$$\lambda_1 = \frac{1}{2}\left[a_{11} + a_{22} + \sqrt{(a_{11} - a_{22})^2 + 4a_{12}^2}\right]$$

$$\lambda_2 = \frac{1}{2}\left[a_{11} + a_{22} - \sqrt{(a_{11} - a_{22})^2 + 4a_{12}^2}\right]$$

[1.20]

由于方程 1.20 根号以下的部分不可能为负，因此，该 (2×2) 对称矩阵的特征值必为实数。

例如，我们有如下矩阵：

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

那么，可以得到：

$$\lambda_1 = \frac{1}{2}\left[1 + 1 + \sqrt{(1-1)^2 + 40.5^2}\right] = 1.5$$

$$\lambda_2 = \frac{1}{2}\left[1 + 1 - \sqrt{(1-1)^2 + 40.5^2}\right] = 0.5$$

要找到特征值为 $\lambda_1 = 1.5$ 时的特征向量，我们需要解齐次方程组，

$$\begin{bmatrix} 1-1.5 & 0.5 \\ 0.5 & 1-1.5 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

得到：

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} x_{21}^* \\ x_{21}^* \end{bmatrix}$$

在这里，任意向量都包含两个相同元。同样，对于 $\lambda_2 = 0.5$，我们要解特征方程组：

$$\begin{bmatrix} 1-0.5 & 0.5 \\ 0.5 & 1-0.5 \end{bmatrix}\begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

得到：

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}\begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} -x_{22}^* \\ x_{22}^* \end{bmatrix}$$

在这里，任意向量都包含两个互为相反数的元。所得特征值下的特征向量可以扩展出一个一维子空间：当指定了特征向量中的一个元时，另一个元也随之可得。可以发现，这里所求的两个特征向量 $\mathbf{x}_1$、$\mathbf{x}_2$ 是互相正交的，即：

$$\mathbf{x}_1 \cdot \mathbf{x}_2 = -x_{21}^* x_{22}^* + x_{21}^* x_{22}^* = 0$$

许多有关 (2×2) 矩阵的特征值和特征向量的性质可以推广到 $(n\times n)$ 矩阵中，尤其是以下几种情况：(1)一个 $(n\times n)$ 矩阵的特征方程 $\det(\mathbf{A}-\lambda\mathbf{I}_n)=0$ 是 λ 的 n 阶多项式，因此，

它的 n 个特征值不一定完全不同[12];(2)矩阵 $\mathbf{A}$ 的所有特征值之和等于 $\mathbf{A}$ 的迹;(3)矩阵 $\mathbf{A}$ 的所有特征值之积等于 $\mathbf{A}$ 的行列式;(4)矩阵 $\mathbf{A}$ 的非零特征值个数等于 $\mathbf{A}$ 的秩;(5)奇异矩阵至少有一个特征值为 0;(6)实对称矩阵的特征值必为实数;(7) 如果矩阵 $\mathbf{A}$ 的所有特征值全都不同(两两均不同),那么特征值下的特征向量可扩张出一个一维子空间;如果有 k 个特征值全相同,那么它们产生的(同一个)特征向量可以扩张出一个 k 维子空间;(8) 不同特征值所产生的特征向量是两两正交的。

假设 $\mathbf{A}$ 为一个$(n\times n)$实对称矩阵,且秩等于 r。让 $\mathbf{\Lambda} \equiv \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_r)$ 表示 $\mathbf{A}$ 的所有非零特征值,$\mathbf{x}_j$ 表示特征值λ_j 下的特征向量,标准化后,我们得到 $\|\mathbf{x}_j\| = 1$。用 $\mathbf{X} \equiv [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_r]$ 表示所有的特征向量,那么,

$$\mathbf{A} = \lambda_1 \mathbf{x}_1 \mathbf{x}_1' + \lambda_2 \mathbf{x}_2 \mathbf{x}_2' + \cdots + \lambda_r \mathbf{x}_r \mathbf{x}_r' = X\Lambda X' \quad [1.21]$$

方程 1.21 称为矩阵 $\mathbf{A}$ 的“谱分解”,它是统计方法中主成分分析和因子分析等方法的基础。

如下为特征值及特征向量的推广运用:假设 $\mathbf{A}$ 为一个 $(n\times n)$ 实对称矩阵,我们可以将方程 1.17 替换成:

$$(\mathbf{A} - \lambda\mathbf{B})\mathbf{x} = \mathbf{0}$$

其中,$\mathbf{B}$ 也是一个 $(n\times n)$ 实对称矩阵,而且是一个正定矩阵。那么,满足该方程的 λ 称为矩阵 $\mathbf{B}$ 下矩阵 $\mathbf{A}$ 的“广义特征值”。我们发现,广义特征值其实是矩阵 $\mathbf{AB}^{-1}$ 的一般特征值。广义特征值和特征向量在多元统计分析中非常有用,如多元线性模型的假设检验。

特征值和特征向量的另一种推广是将其运用在长方形

矩阵中。假设矩阵 $\mathbf{A}$ 为 $(m \times n)$ 矩阵，且其秩为 r。那么，$\mathbf{A}$ 可分解为：

$$\mathbf{A} = \underset{(m\times m)}{\mathbf{B}} \begin{bmatrix} \underset{(r\times r)}{\mathbf{\Lambda}} & \underset{(r\times n-r)}{\mathbf{0}} \\ \underset{(m-r\times r)}{\mathbf{0}} & \underset{(m-r\times n-r)}{\mathbf{0}} \end{bmatrix} \underset{(n\times n)}{\mathbf{C}'} \qquad [1.22]$$

其中，(1) 矩阵 $\mathbf{B}$ 和矩阵 $\mathbf{C}$ 为正交矩阵，但不唯一；(2) $\mathbf{\Lambda}^2$ 是一个对角矩阵，它包含矩阵 $\mathbf{A}'\mathbf{A}$ 和 $\mathbf{A}\mathbf{A}'$（其所包含的特征值相同）的所有非零特征值；(3) 并不是有所有零矩阵都会用到（当然，如果 $r=m=n$，那么方程 1.22 可以简化为方程 1.21 的谱分解）。

方程 1.22 称做矩阵 $\mathbf{A}$ 的“奇异值分解”，矩阵 $\mathbf{\Lambda}$ 的对角元为矩阵 $\mathbf{A}$ 的奇异值（因此是矩阵 $\mathbf{A}'\mathbf{A}$ 和 $\mathbf{A}\mathbf{A}'$特征值的平方根）。奇异值分解非常有用，比如在提高最小二乘计算的效率和精度上。

第6节 | 二次型及正定矩阵

表达式

$$\underset{(1\times n)}{\mathbf{x}'}\underset{(n\times n)}{\mathbf{A}}\underset{(n\times 1)}{\mathbf{x}} \qquad [1.23]$$

称为“x的二次型”。在本节里,矩阵A从始至终表示一个实对称矩阵。如果对于所有非负x表达式1.23都为正,那么,我们说矩阵A为正定矩阵。对于所有非负x,表达式1.23都为非负(即正或0),那么,矩阵A为半正定矩阵。一个正定矩阵的所有特征值均为正(因此,正定矩阵是非奇异矩阵),一个半正定矩阵的所有特征值均为正或者为0。

请看以下方程:

$$\underset{(m\times m)}{\mathbf{C}} = \underset{(m\times n)}{\mathbf{B}'}\underset{(n\times n)}{\mathbf{A}}\underset{(n\times m)}{\mathbf{B}}$$

其中,A为正定矩阵,B为列满秩矩阵,$m \leqslant n$。我会证明,矩阵C同样是正定矩阵。注意,首先矩阵C是对称的。

$$\mathbf{C}' = (\mathbf{B}'\mathbf{A}\mathbf{B})' = \mathbf{B}'\mathbf{A}'\mathbf{B} = \mathbf{B}'\mathbf{A}\mathbf{B} = \mathbf{C}$$

如果y是任意$(m\times 1)$非零向量,那么$\underset{(n\times 1)}{\mathbf{x}} = \mathbf{By}$也为非零向量。因为矩阵B的秩为$m$,我们可以从B中选择$m$个线性独立的行组成一个非奇异矩阵$\mathbf{B}^*$。那么,$\underset{(n\times 1)}{\mathbf{x}^*} = \mathbf{B}^*\mathbf{y}$,它包括向量x中所包含元的子集,且也是非零的,原因在于$\mathbf{y} =$

$\mathbf{B}^{*-1}\mathbf{x}^* \neq \mathbf{0}$。因此，$\mathbf{y}'\mathbf{C}\mathbf{y}=\mathbf{y}'\mathbf{B}'\mathbf{A}\mathbf{B}\mathbf{y}=\mathbf{x}'\mathbf{A}\mathbf{x}$ 必然为正，所以矩阵 $\mathbf{C}$ 为正定矩阵。同理，如果 $\text{rank}(\mathbf{B}) < m$，那么矩阵 $\mathbf{C}$ 为半正定矩阵。如果 $\mathbf{B}$ 为列满秩矩阵，那么矩阵 $\underset{(m\times n)}{\mathbf{B}'}\underset{(n\times m)}{\mathbf{B}} = \mathbf{B}'\mathbf{I}_n\mathbf{B}$ 为一个正定矩阵(因为矩阵 $\mathbf{I}_n$ 明显为一个正定矩阵)，否则为半正定矩阵。

正定矩阵和半正定矩阵，如方差—协方差矩阵、相关矩阵平方和和乘积矩阵，在统计中都起着至关重要的作用。

Cholesky 分解

每个对称正定矩阵 $\mathbf{A}$ 都可以被唯一地写为 $\mathbf{A}=\mathbf{U}'\mathbf{U}$，其中，$\mathbf{U}$ 是一个对角元素为正的上三角矩阵。$\mathbf{U}$ 称为矩阵 $\mathbf{A}$ 的"Cholesky 因子"，或者可以看成是某种矩阵的平方根。

现在我们来考虑一个 (3×3) 矩阵：

$$\mathbf{A}=\begin{bmatrix}1.0 & 0.5 & 0.3\\ 0.5 & 1.0 & 0.5\\ 0.3 & 0.5 & 1.0\end{bmatrix}$$

同时用矩阵 $\mathbf{U}$

$$\mathbf{U}=\begin{bmatrix}u_{11} & u_{12} & u_{13}\\ 0 & u_{22} & u_{23}\\ 0 & 0 & u_{33}\end{bmatrix}$$

来表示矩阵 $\mathbf{A}$ 的 Cholesky 因子。那么，

$$\mathbf{U}'\mathbf{U}=\begin{bmatrix}u_{11}^2 & u_{11}u_{12} & u_{11}u_{13}\\ u_{12}u_{11} & u_{12}^2+u_{22}^2 & u_{12}u_{13}+u_{22}u_{23}\\ u_{13}u_{11} & u_{13}u_{12}+u_{23}u_{22} & u_{13}^2+u_{23}^2+u_{33}^2\end{bmatrix}$$

$$= \begin{bmatrix} 1.0 & 0.5 & 0.3 \\ 0.5 & 1.0 & 0.5 \\ 0.3 & 0.5 & 1.0 \end{bmatrix} = \mathbf{A}$$

进而得到：

$$u_{11}^2 = 1.0 \rightarrow u_{11} = 1.0$$

$$u_{12}u_{11} = u_{12} \times 1 = 0.5 \rightarrow u_{12} = 0.5$$

$$u_{12}^2 + u_{22}^2 = 0.5^2 + u_{22}^2 = 1 \rightarrow u_{22} = \sqrt{1-0.5^2} = 0.8660$$

$$u_{13}u_{11} = u_{13} \times 1 = 0.3 \rightarrow u_{13} = 0.3$$

$$u_{13}u_{12} + u_{23}u_{22} = 0.3 \times 0.5 + u_{23} \times 0.8660 = 0.5 \rightarrow$$

$$u_{23} = (0.5 - 0.3 \times 0.5)/0.8660 = 0.4041$$

$$u_{13}^2 + u_{23}^2 + u_{33}^2 = 0.3^2 + 0.4141^2 + u_{33}^2 = 1 \rightarrow$$

$$u_{33} = \sqrt{1 - 0.3^2 - 0.4141^2} = 0.8641$$

因此，

$$\mathbf{U} = \begin{bmatrix} 1.0 & 0.5 & 0.3 \\ 0 & 0.8660 & 0.4041 \\ 0 & 0 & 0.8641 \end{bmatrix}$$

这个过程可以引申到任意秩的对称正定矩阵上。[13]

第 7 节 | 推荐阅读

有关矩阵及线性代数的书籍很多，大多数仅仅描述了有关向量空间的基本属性，却没有提供详细的图解。

关于矩阵的书籍，包括希利(Healy, 1986)、格雷比尔(Graybill, 1983)、瑟尔(Searle, 1982)以及格林(Green)和卡罗尔(Carroll)(1976)的研究，均主要针对统计应用。后几本的几何描述很详细。

戴维斯(Davis, 1973)的著作对矩阵代数的描述清晰且简单，包括一些向量几何内容，但较为有限，仅局限于二维空间。

南布狄瑞(Namboodiri, 1984)的著作关注矩阵代数的解释，结构紧凑，但是不包括向量几何。

有关统计计算的书籍，有肯尼迪(Kennedy)和金特尔(Gentle)(1980)及莫纳汉(Monahan)(2001)等人的著作，主要描述了矩阵和线性代数在数字计算机中的应用。

第2章

微积分入门

微积分主要处理两种问题：寻找曲线的切线斜率（微分）和计算曲线下方的面积（积分）。早在17世纪，英国物理学家、数学家艾萨克·牛顿爵士（Sir Isaac Newton）和德国数学家、哲学家戈特弗里德·威廉·凡·莱布尼茨 Gottfriend Wilhelm von Leibniz）就各自独立地证明了这两种问题的联系，进一步巩固并发展了古典时代的数学。因此，牛顿和莱布尼茨是公认的微积分创始人。[14]到了19世纪，伟大的法国数学家奥古斯丁·路易斯·柯西（Augustin Louis Cauchy）与其他学者一起引入了极限的概念，从而为微积分建立了一个在逻辑上更为严格的基础。

在本章中，我们首先简单回顾一些基础数学，然后，按如下次序简要地介绍微积分：方程的极限；方程的求导；利用求导解决最优化问题；多变量的偏导、条件最优化和矩阵的微积分；泰勒展式和渐近式；积分学的重要思想。

虽然我的叙述远不够严格、透彻，但是读者仍可以从中获得许多对微积分基本问题的直观认识。

第1节 | 回顾

数字

对不同类别数字的定义取决于所要研究数学问题的深度，对于社会科学，如下基本定义已经可以基本满足我们的研究需要了：

第一，自然数包括0及所有正整数。[15]

第二，整数包括所有负整数、正整数和0。

第三，整数和分数统称为有理数，任何一个有理数都可以写成分数$\frac{n}{m}$（n和m是整数，且$m \neq 0$）的形式。如$-\frac{1}{2}$和$\frac{123}{4}$。

第四，实数包括所有的有理数和无理数，例如，$\sqrt{2} \approx 1.41421$，数学常数$\pi \approx 3.14159$及$e \approx 2.71828$，这些数都不能写成两个整数的比例。所有实数可以投影到一条连续的直线上，从$-\infty$到$+\infty$。

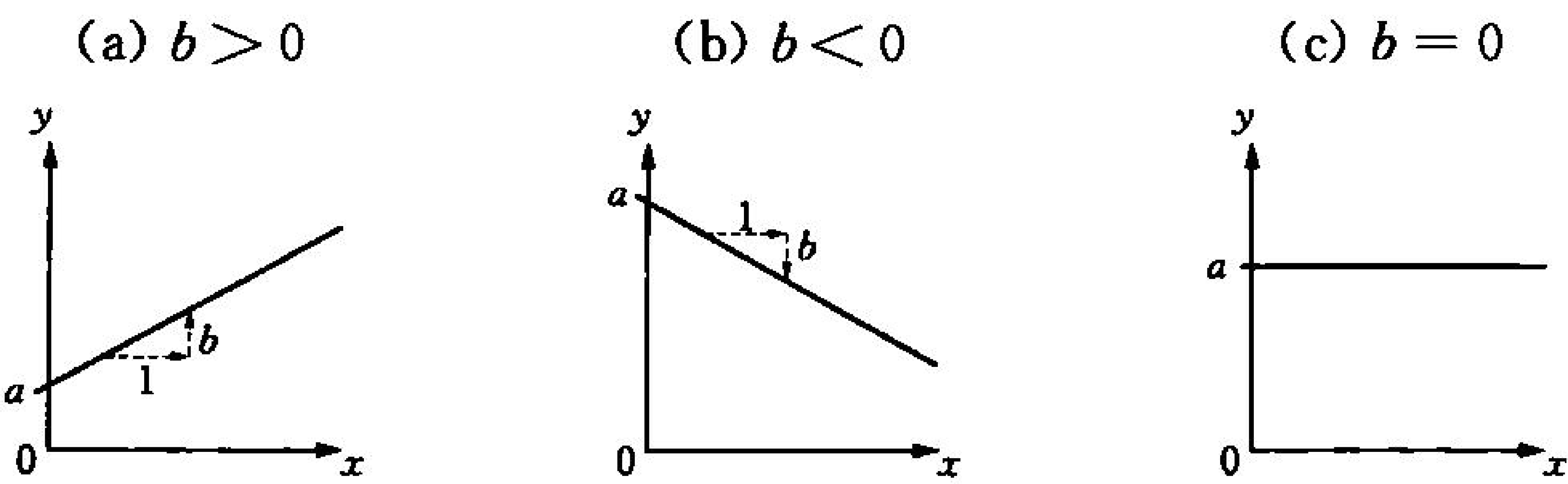
第五，复数可以用$a+bi$表示，其中，a和b是实数，i是虚数，$i \equiv \sqrt{-1}$。在直角坐标系中，复数可以想象成复平面上的点——横轴即实轴对应于实数部分a，纵轴即虚轴对应于虚数部分的系数bi。当$b=0$时，复数即实数。

线和平面

直线可以用方程表示：

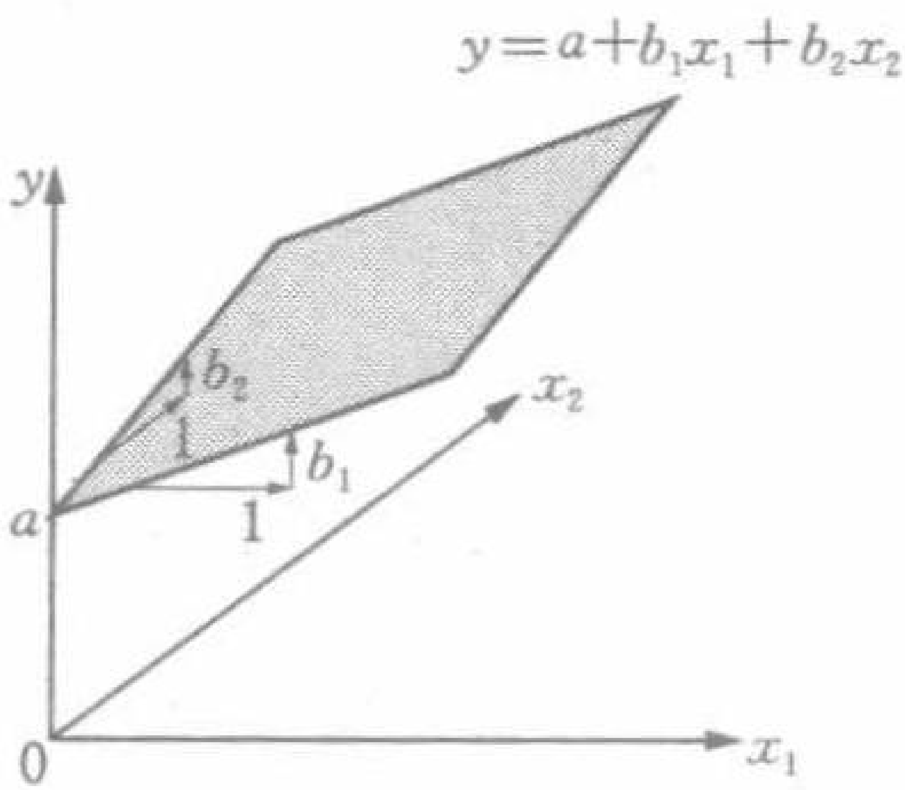
$$y = a + bx$$

其中，a 和 b 是常数，且 a 是 y 轴截距（$x = 0$ 时的 y 值），b 是斜率（x 增加 1 时 y 的变化）。图 2.1 表示在二维坐标下以 x 和 y 为轴的直线，对于每一种情况，直线都是可以向左右无限延伸的。如果斜率是正的（$b > 0$），直线从西南往东北延伸；如果斜率是负的（$b < 0$），直线从西北往东南延伸；如果 $b = 0$，直线是水平的。



注：(a) $b > 0$；(b) $b < 0$；(c) $b = 0$。

图 2.1 直线 $y = a + bx$ 的图像



注：这里斜率 b_1、b_2 都是正值。

图 2.2 平面方程 $y=a+b_1x_1+b_2x_2$

同样，我们有线性方程：

$$y = a + b_1x_1 + b_2x_2$$

它代表三维空间的一个以 x_1、x_2 和 y 为轴的平面，如图 2.2 所示。x_1、x_2 和 y 轴两两垂直，我们可以把 x_2 轴的方向想

象成垂直于纸面向内，且平面在各个方向无限延伸。据图，截距 a 表示在 x_1 和 x_2 都为 0 时的 y 值；b_1 表示固定了 x_2 值后，平面在 x_1 方向上的斜率；b_2 表示固定 x_1 值后，平面在 x_2 方向上的斜率。

直线方程还可以表示为其他形式：

$$cx + dy = e$$

将其转换为截距式为：

$$y = \frac{e}{d} - \frac{c}{d}x$$

同样，方程

$$c_1x_1 + c_2x_2 + dy = e$$

可以表示平面

$$y = \frac{e}{d} - \frac{c_1}{d}x_1 - \frac{c_2}{d}x_2$$

多项式

多项式具有以下形式：

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_px^p$$

其中，a_0，a_1，a_2，…，a_p 是常数，除 a_p 外，其他系数可为 0。最大的指数 p 为多项式的阶。如图 2.3 所示，一阶多项式即一条直线：

$$y = a_0 + a_1x$$

二阶多项式是二次方程：

$$y = a_0 + a_1x + a_2x^2$$

三阶多项式是三次方程：

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

一个 p 阶的多项式有 $p-1$ 个弯。例如，二阶多项式有一个弯，三阶多项式有两个弯，等等。

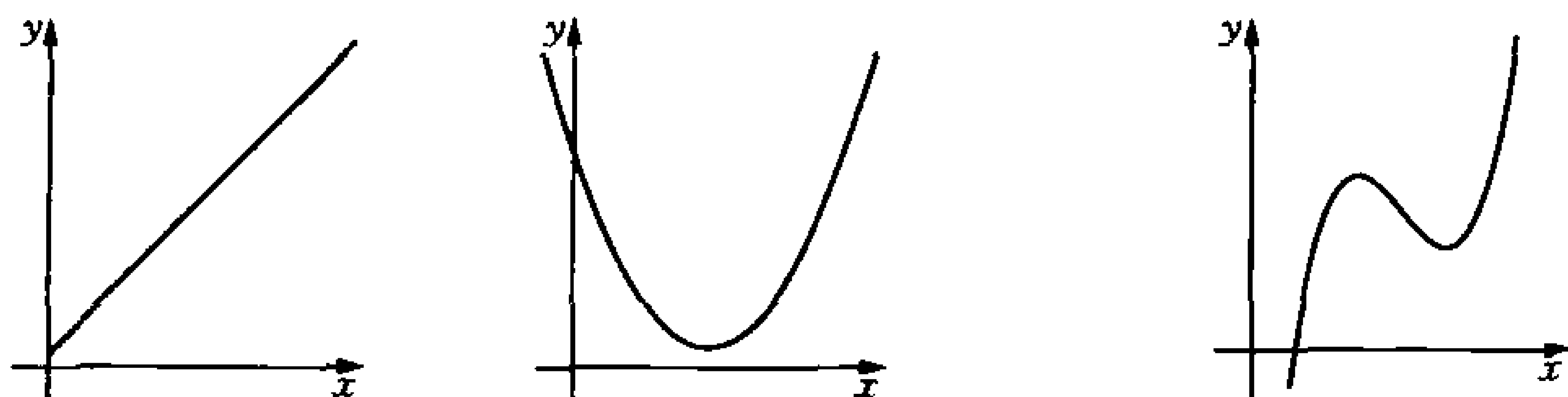


图 2.3 "典型"的一阶(线性)、二阶(二次型)、三阶(三次方)多项式

指数和对数

对数方程：

$$\log_b x = y$$

读作"以 b 为底 x 为真数的对数是 y"，其等价于

$$x = b^y$$

其中，$b > 0$ 且 $b \neq 1$。

$$\log_{10} 10 = 1 \quad 因为\ 10^1 = 10$$
$$\log_{10} 100 = 2 \quad 因为\ 10^2 = 100$$
$$\log_{10} 1 = 0 \quad 因为\ 10^0 = 1$$
$$\log_{10} 0.1 = -1 \quad 因为\ 10^{-1} = 0.1$$

同样，

$$\log_2 2 = 1 \quad 因为\ 2^1 = 2$$

$$\log_2 4 = 2 \qquad 因为\ 2^2 = 4$$

$$\log_2 1 = 0 \qquad 因为\ 2^0 = 1$$

$$\log_2 \frac{1}{4} = -2 \qquad 因为\ 2^{-2} = \frac{1}{4}$$

实际上，不论底为何数，只要真数为 1，其对数就为 0，因为 $b^0 = 1(b \neq 0)$。在对数函数中，x 的定义域为 $x > 0$。数学中有一些常用的底，如数学常数 $e \approx 2.71828$，其中，以 e 为底的对数都称为“自然对数”。[16]

典型的对数方程不管其底如何，都具有类似的形状，如图 2.4 所示。有时为方便计算，我们常常需要将对数函数的底换为另一个常数或字符，这时，我们得到：

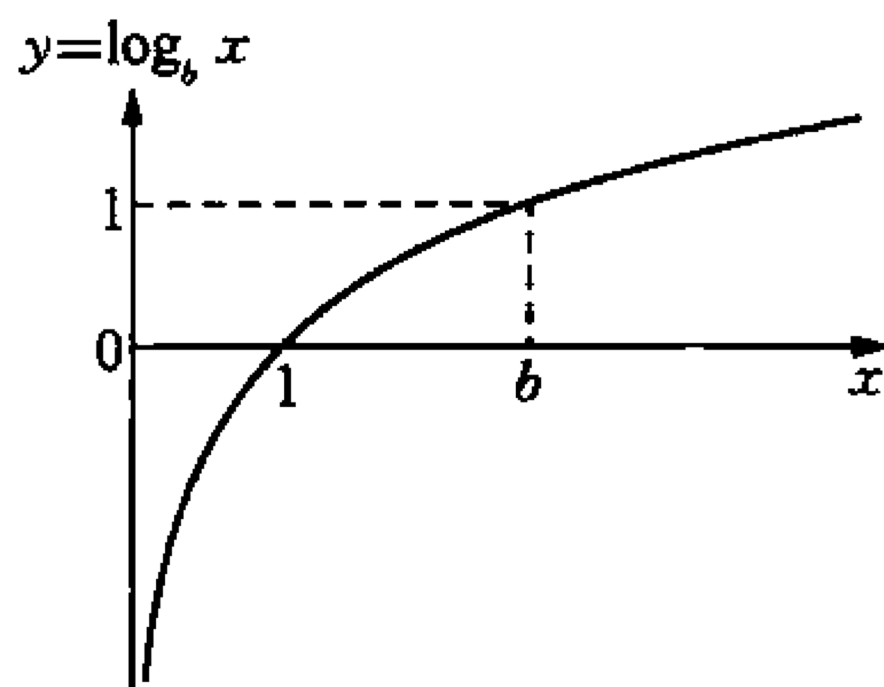


图 2.4 对数函数 $y = \log_b x$

$$\log_a x = \log_a b \times \log_b x$$

该公式为换底公式。例如，

$$\log_{10} 1000 = 3 = \log_{10} 2 \times \log_2 1000 \approx 0.301030 \times 9.965784$$

对数继承了指数的一些特性，如 $b^{x_1} b^{x_2} = b^{x_1+x_2}$，因此，

$$\log(x_1 x_2) = \log x_1 + \log x_2$$

同样，$\frac{b^{x_1}}{b^{x_2}} = b^{x_1-x_2}$，因此：

$$\log\left(\frac{x_1}{x_2}\right) = \log x_1 - \log x_2$$

$b^{ax} = (b^x)^a$，那么，

$$\log(x^a) = a\log x$$

为了简化繁冗的计算，我们曾将乘法转化为加法、除法转化

为减法、指数转化为乘法。虽然现在已经不需要这么做了，但对数仍然在数学及统计学中扮演着不可或缺的角色。

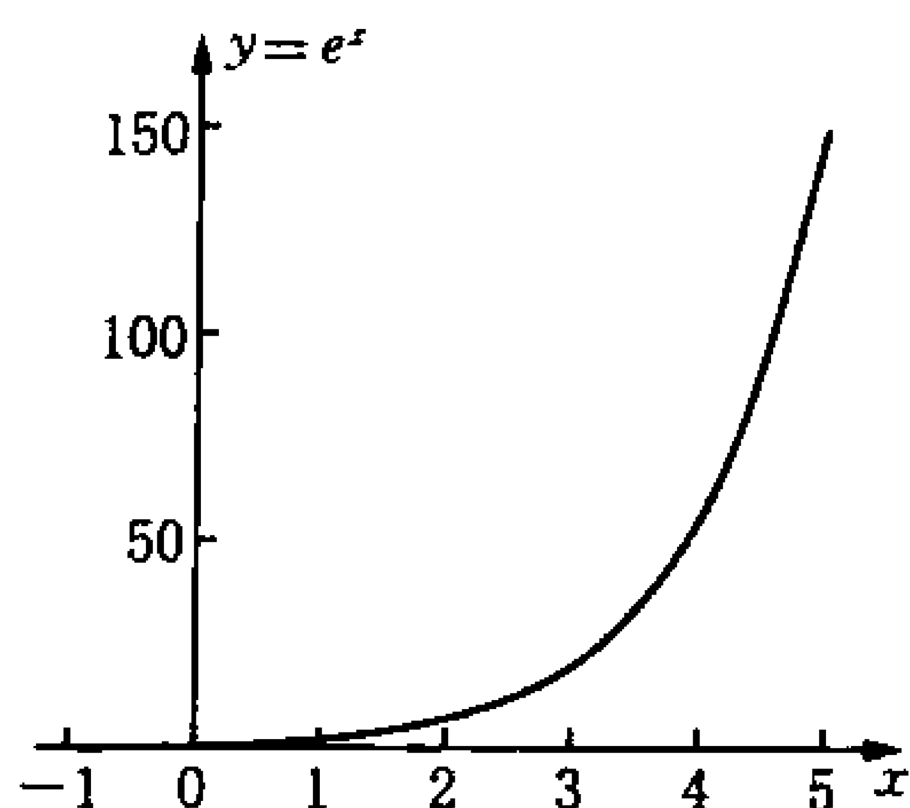


图 2.5 指数函数 $y=e^x$

指数方程具有这样的形式：

$$y=a^x$$

其中，a 是常数。常用的指数有 $y=\exp(x)=e^x$，如图 2.5 所示。对数函数和指数函数互为反函数：$\log_a(a^x)=x$，$a^{\log_a x}=x$。

三角函数

图 2.6 为一个单位圆——一个圆心在原点，半径为 1 的圆。角 x 在圆内生成了一个直角三角形，同时，该夹角是以水平轴为起始轴，按逆时针方向旋转测量得出的。

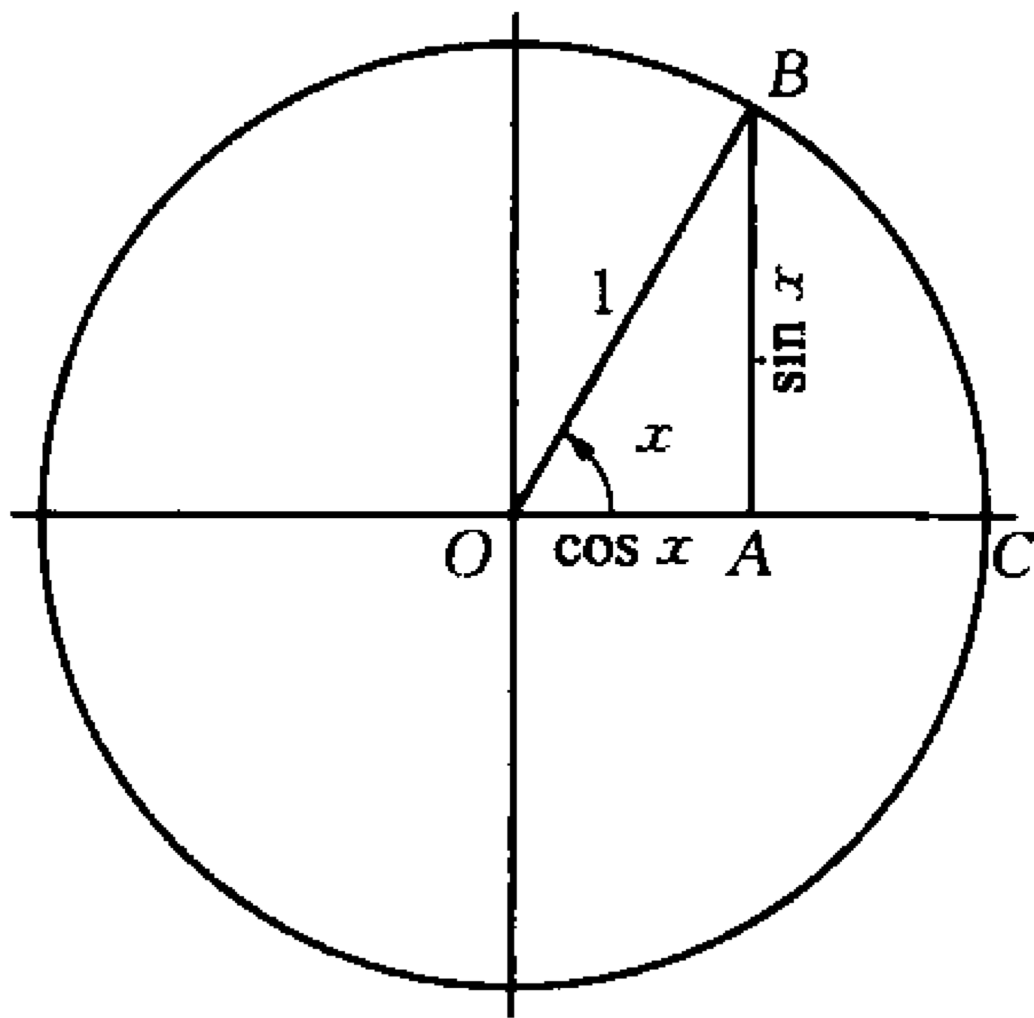


图 2.6 单位圆、夹角及其正弦和余弦

角 x 的余弦即邻边/斜边(OA/OB)，记为 $\cos x$，长度等于 OA(因为 $OB=1$)；角 x 的正弦即对边/斜边(AB/OB)，记为 $\sin x$，长度等于 AB；角 x 的正切即对边/邻边(AB/OA)，记为 $\tan x=\sin x/\cos x$。

如图 2.7 所示，正弦、余弦、正切函数角的取值范围为 $-360°$ 到 $360°$，其中负值表示顺时针方向旋转得到的角。当夹角趋近于 $\pm 90°$ 或 $\pm 270°$ 时，正切函数值相应地趋近于 $\pm\infty$。

此外，$\sin x = \cos(x-90)$。

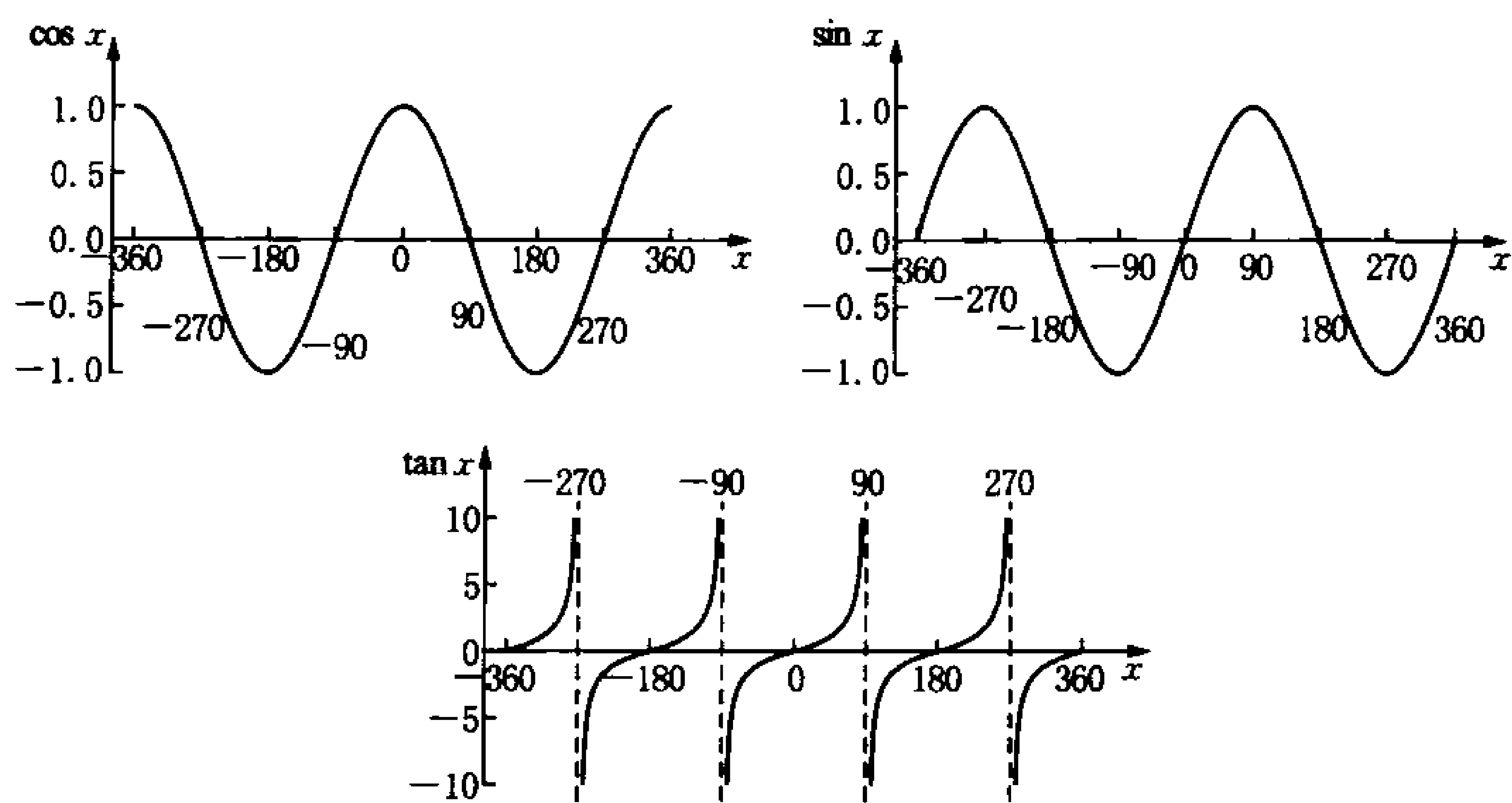


图 2.7　角度在 $x=-360°$ 和 $x=360°$ 之间的正弦、余弦和正切函数

有时，用弧度来度量角会更加方便，2π 弧度等价于 360° 角度。如图 2.6，由于单位圆的周长为 2π，因此我们可以用夹角两边所夹的弧长来代表弧度，如角 x 的弧度为 BC。

第2节 | 极限

微积分常用来处理具有 $y=f(x)$ 形式的函数，我们所考虑的定义域（自变量的取值）和值域（因变量的取值）都是实数。极限用于考虑函数在其自变量 x 趋近但不等于某个数值时的行为。这是一个很重要的思想，尤其在函数没有对自变量 x 的某些数值给出定义或者函数在某些数值中没有意义的情况下。

极限的"ε—δ"定义

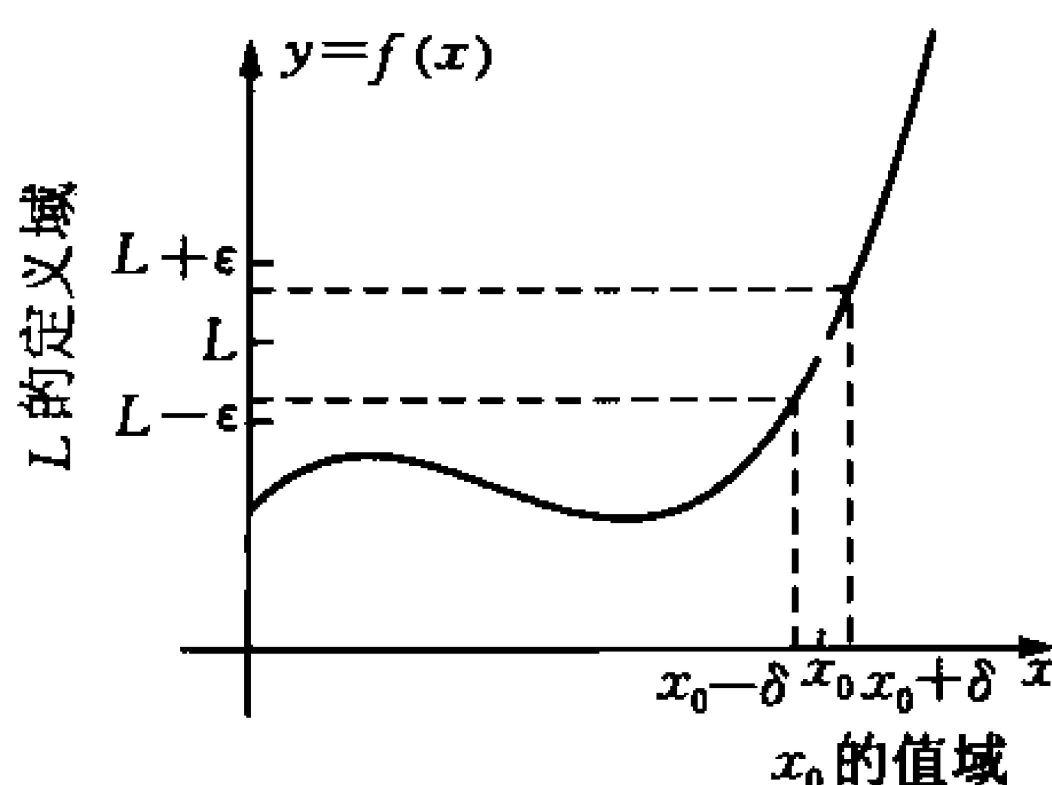
函数 $y=f(x)$ 在 $x=x_0$（即一个特定的 x 值）处有极限 L。对于任意正数 ε，无论它多小，总是存在一个正数 δ，只要满足 x 与 x_0 的距离小于 δ，即只要 x 位于 x_0 左右两边足够小的 2δ 值域中，$f(x)$ 和 L 的距离即小于 ε。用符号表示为：

$$|f(x)-L|<\varepsilon \quad (0<|x-x_0|<\delta)$$

图2.8形象地描述了这一定义。注意，其中 $f(x_0)$ 不需要等于 L。实际上，极限函数当 $f(x)$ 在 $x=x_0$ 不存在的时候最有用。若 L 是 $f(x)$ 在 $x=x_0$ 时的极限，那么，这意味着当 x 从 x_0 左右两边趋近 x_0 时，$f(x)$ 趋近于 L。用公式表示为：

$$\lim_{x \to x_0} f(x) = L$$

读作“$f(x)$在 x 趋近于 x_0 时的极限为 L”。



注：x_0 上方对应的曲线的缺口表示当 $x = x_0$ 时函数值无法定义。

图 2.8 $\lim_{x \to x_0} f(x) = L$：函数 $f(x)$的极限

一个例子：如何找极限

让我们来找出函数 $y = f(x) = \dfrac{x^2 - 1}{x - 1}$ 在 $x_0 = 1$ 时的极限。我们发现，$f(1) = \dfrac{1-1}{1-1} = \dfrac{0}{0}$ 是没有意义的（分母为 0）。尽管如此，只要 x 不等于 1，无论它多么接近 1，我们都可以将等式上下除以 $x - 1$：

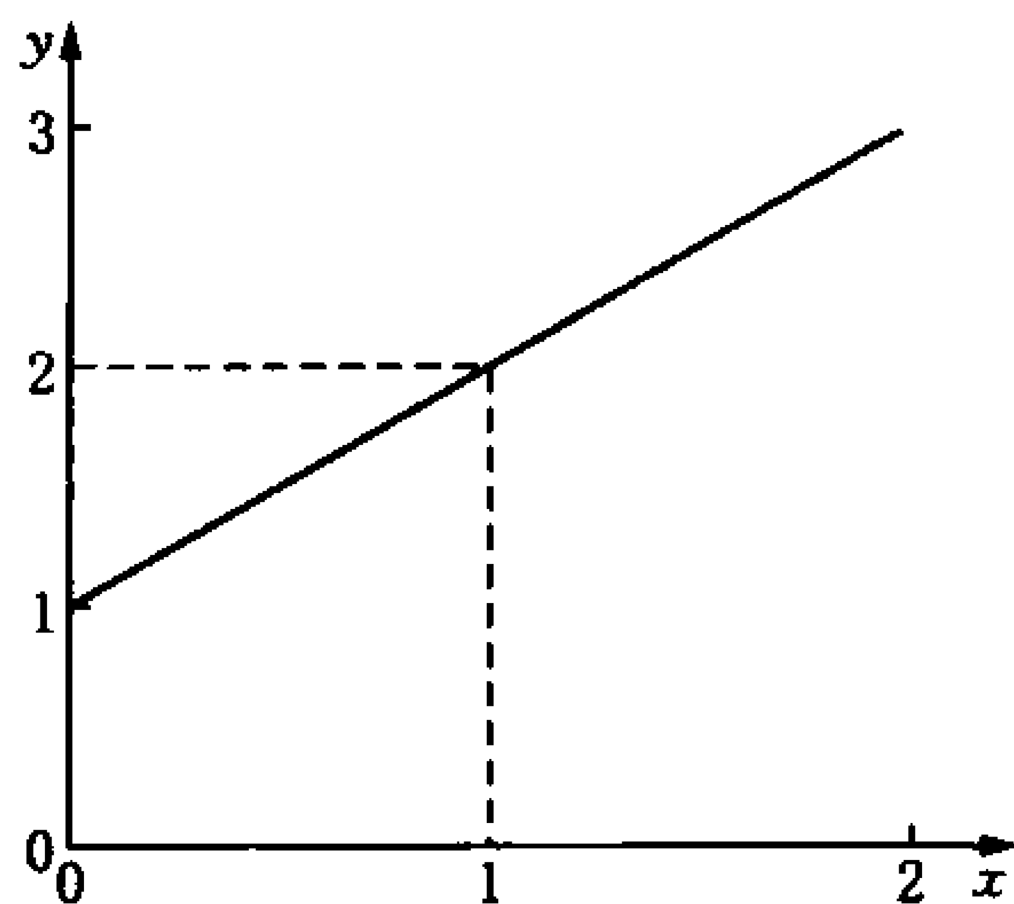


图 2.9 $\lim_{x \to 1} \dfrac{x^2 - 1}{x - 1} = 2 (x \neq 1)$

$$y = \frac{x^2 - 1}{x - 1} = \frac{(x+1)(x-1)}{x-1} = x + 1$$

因为 $x_0 + 1 = 1 + 1 = 2$，所以

$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = \lim_{x \to 1}(x + 1) = 1 + 1 = 2$$

图 2.9 展示了这个极限。

极限运算规则

假设 $f(x)$和 $g(x)$是自变量 x 的两个函数，且在 $x=x_0$ 时都有极限：

$$\lim_{x \to x_0} f(x) = a$$

$$\lim_{x \to x_0} g(x) = b$$

那么，$f(x)$和 $g(x)$的极限的算术运算如下：

$$\lim_{x \to x_0}[f(x) + g(x)] = a + b$$

$$\lim_{x \to x_0}[f(x) - g(x)] = a - b$$

$$\lim_{x \to x_0}[f(x)g(x)] = ab$$

$$\lim_{x \to x_0}[f(x)/g(x)] = a/b \quad (b \neq 0)$$

同样，假设 c 和 n 是常数，且 $\lim_{x \to x_0} f(x) = a$，那么，

$$\lim_{x \to x_0} c = c$$

$$\lim_{x \to x_0}[cf(x)] = ca$$

$$\lim_{x \to x_0}\left\{[f(x)]^n\right\} = a^n$$

因此，

$$\lim_{x \to x_0} x = x_0$$

第3节 | 函数求导

现在考虑函数 $y = f(x)$ 在 x 的两个值下的情况：

$$x = x_1, \ y_1 = f(x_1)$$

$$x = x_2, \ y_2 = f(x_2)$$

差商是指从点(x_1, y_1)到点(x_2, y_2)时，y 值的变化除以 x 值的变化，即：

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

其中，Δ 读作“Delta”，是“变化”的简写。如图 2.10 所示，差商是连接点(x_1, y_1)和点(x_2, y_2)的割线。

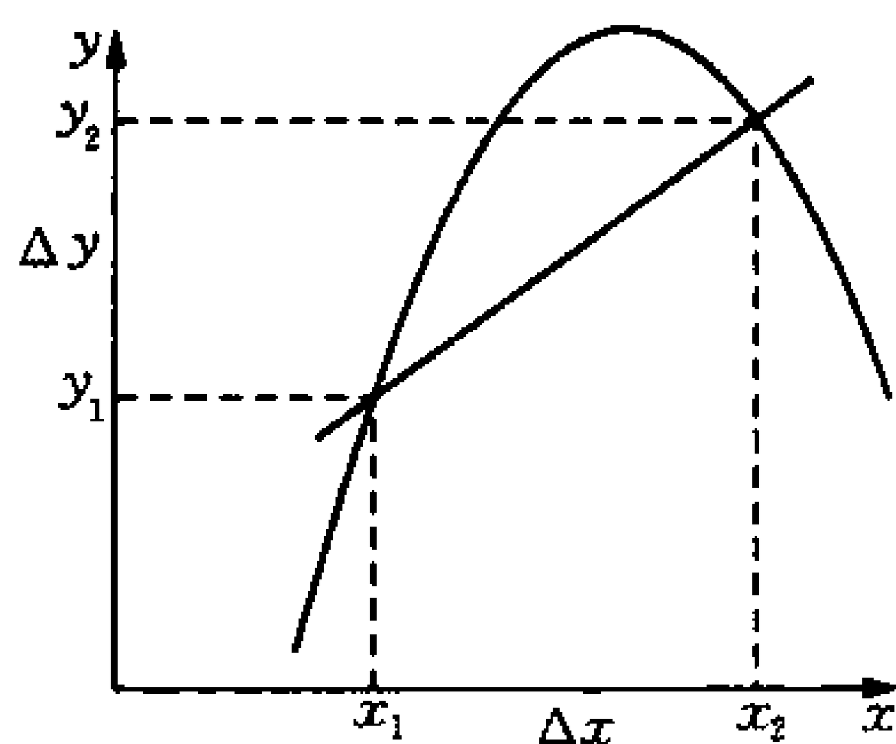


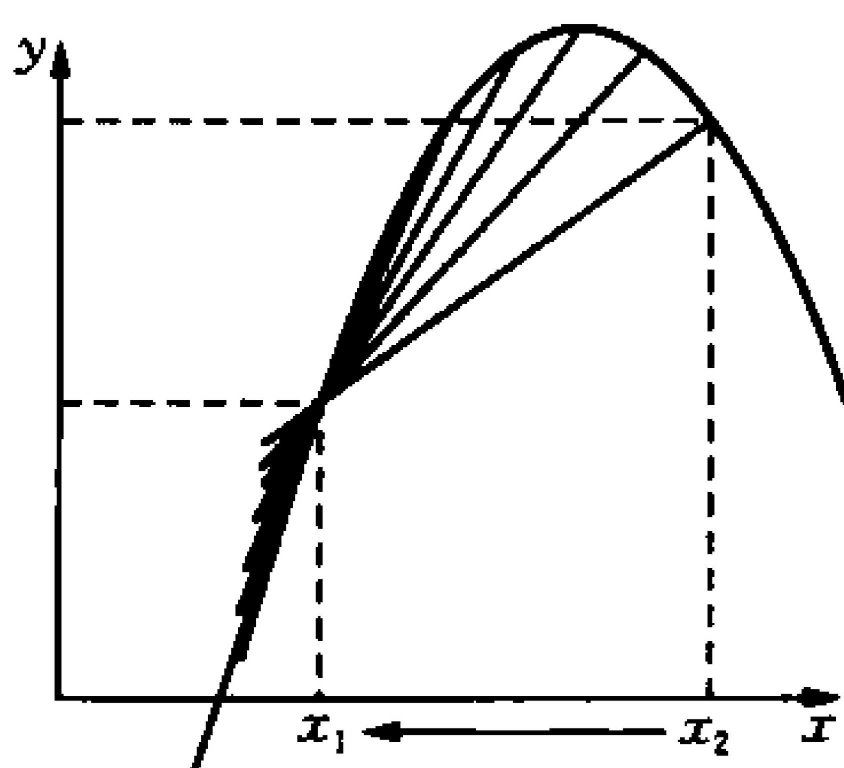
图 2.10 差商 Δy/Δx 是连接(x_1, y_1)和(x_2, y_2)两点的割线的斜率

$f(x)$在 $x = x_1$ 时的导数是差商$\frac{\Delta y}{\Delta x}$在 x_2 接近 x_1 时的极限(即 $\Delta x \to 0$)：

$$\frac{dy}{dx} = \lim_{x_2 \to x_1} \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

$$= \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x) - f(x_1)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

如图 2.11 所示，导数是 $f(x)$在 $x = x_1$ 时的切线。



注：随着 x_2 逐渐趋近于 x_1，割线越来越趋近于切线。

图 2.11　导数是 $f(x_1)$的切线斜率

我们还可以用下面的符号表示导数：

$$\frac{dy}{dx} = \frac{df(x)}{dx} = f'(x)$$

表达式 $f'(x)$强调了导数是 x 本身的一个函数。对于 dx 和 dy，可以将其想象成无限小但是不等于 0 的数值，在很多情况下的求导可以把它们当做数字来处理。求函数导数的过程叫做“微分”。

导数——差商的极限

给定函数 $y = f(x) = x^2$，求任意 x 的 $f'(x)$。

运用导数是差商的极限的定义：

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{(x + \Delta x)^2 - x^2}{\Delta x}$$

$$= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + \Delta x^2 - x^2}{\Delta x} = \lim_{\Delta x \to 0} \frac{2x\Delta x + \Delta x^2}{\Delta x}$$

$$= \lim_{\Delta x \to 0}(2x + \Delta x) = \lim_{\Delta x \to 0} 2x + \lim_{\Delta x \to 0} \Delta x$$

$$= 2x + 0 = 2x$$

由于 Δx 虽然接近于 0，但是永远不等于 0，因此除法是合适的。例如，曲线 $y = f(x) = x^2$ 在 $x = 3$ 时的切线是 $f'(x) = 2x = 2 \times 3 = 6$。

幂函数的导数

一般而言，$y = f(x) = ax^n$ 的导数为：

$$\frac{dy}{dx} = nax^{n-1}$$

例如，$y = 3x^6$ 的导数是：

$$\frac{dy}{dx} = 6 \times 3x^{6-1} = 18x^5$$

负幂函数和分数幂函数的导数是类似的，例如，$y = \frac{1}{4x^3} = \frac{1}{4}x^{-3}$ 的导数是：

$$\frac{dy}{dx} = -3 \times \frac{1}{4}x^{-3-1} = -\frac{3}{4}x^{-4} = -\frac{3}{4x^4}$$

$y = \sqrt{x} = x^{\frac{1}{2}}$ 的导数是：

$$\frac{dy}{dx} = \frac{1}{2}x^{\frac{1}{2}-1} = \frac{1}{2}x^{-\frac{1}{2}} = \frac{1}{2\sqrt{x}}$$

导数的运算规则

假设一个函数是另外两个函数的和：

$$h(x) = f(x) + g(x)$$

导数的加法法则与极限函数的加法规则一样，为 $h'(x) = f'(x) + g'(x)$。例如，

$$y = 2x^2 + 3x + 4$$

$$\frac{dy}{dx} = 4x + 3 + 0 = 4x + 3$$

注意，常数的导数（如上例中常数 4 的导数）为 0，因为常数可以表示为：

$$y = f(x) = 4 = 4x^0$$

该结果的几何意义是，一个常数可以用 $\{x, y\}$ 平面的一条水平直线表示，而这条直线的斜率为 0。

对于多项式函数，导数的加法规则同样适用：

$$\frac{d}{dx}ax^n = nax^{n-1}$$

导数的乘法和除法规则比较复杂。导数的乘法规则为：

$$h(x) = f(x)g(x)$$

$$h'(x) = f(x)g'(x) + f'(x)g(x)$$

导数的除法规则为：

$$h(x) = f(x)/g(x)$$

$$h'(x)=\frac{g(x)f'(x)-g'(x)f(x)}{[g(x)]^2}$$

例如，$y=(x^2+1)(2x^3-3x)$ 的导数为：

$$\frac{dy}{dx}=(x^2+1)(6x^2-3)+2x(2x^3-3x)$$

又如，$y=\frac{x}{x^2-3x+5}$ 的导数为：

$$\frac{dy}{dx}=\frac{x^2-3x+5-(2x-3)x}{(x^2-3x+5)^2}=\frac{-x^2+5}{(x^2-3x+5)^2}$$

导数的链式法则为，假设 y 是 x 的间接函数，$y=f(z)$，$z=g(x)$：

$$y=f[g(x)]=h(x)$$

那么，y 关于 x 的导数为：

$$h'(x)=\frac{dy}{dx}=\frac{dy}{dz}\times\frac{dz}{dx}$$

看上去分子和分母中的导数 dz 可以消去。[17]

例如，求函数 $y=(x^2+3x+6)^5$ 中 y 关于 x 的导数 $\frac{dy}{dx}$。我们可以展开幂函数(即括号里的表达式乘以它自身五次)，但是这样会使运算极其复杂。如果我们运用链式法则，会使运算简单得多。首先，引入一个新的变量 z，代表括号里的表达式：

$$z=g(x)=x^2+3x+6$$

那么，

$$y=f(z)=z^5$$

然后用 y 对 z，z 对 x 分别求导数得到：

$$\frac{dy}{dz}=5z^4$$

$$\frac{dz}{dx}=2x+3$$

运用链式法则得到：

$$\frac{dy}{dx}=\frac{dy}{dz}\times\frac{dz}{dx}=5z^4(2x+3)$$

最后，用 x 替代 z，得到：

$$\frac{dy}{dx}=5(x^2+3x+6)^4(2x+3)$$

此例是典型的链式法则运用：引入一个“人为”的变量来简化表达式的结构。

指数函数和对数函数的导数

在应用统计中，我们经常会碰到指数函数和对数函数，因此，知道如何求这些函数的导数是很重要的。

对数函数 $y=\log_e(x)$ 的导数是：

$$\frac{d\log_e(x)}{dx}=\frac{1}{x}=x^{-1}$$

其中，$\log_e$ 是自然对数，即以 $e\approx 2.71828$ 为底的对数。

事实上，简单的导数形式是自然对数称为“自然”的原因之一。

指数函数 $y=e^x$ 的导数为：

$$\frac{de^x}{dx} = e^x$$

对于任意常数 a 的指数函数 $y = a^x$，其导数为：

$$\frac{da^x}{dx} = a^x \log_e a$$

三角函数的导数

基本三角函数的导数如下，其中，x 是以弧度为单位的：

$$\frac{d\cos x}{dx} = -\sin x$$

$$\frac{d\sin x}{dx} = \cos x$$

$$\frac{d\tan x}{dx} = \frac{1}{\cos^2 x} \left(x \neq \frac{\pi}{2}, \pm\frac{3\pi}{2}, \text{即} \cos x \neq 0\right)$$

二阶或高阶导数

因为导数是它本身的函数，所以可以被再次求导。函数 $y = f(x)$ 的二阶导数为：

$$f''(x) = \frac{d^2 y}{dx^2} = \frac{df'(x)}{dx}$$

同样，$y = f(x)$ 的三阶导数是二阶导数的导数：

$$f'''(x) = \frac{d^3 y}{dx^3} = \frac{df''(x)}{dx}$$

高阶导数以此类推。

例如，函数 $y = f(x) = 5x^4 + 3x^2 + 6$ 的各阶导数为：

$$f'(x) = 20x^3 + 6x$$

$$f''(x) = 60x^2 + 6$$

$$f'''(x) = 120x$$

$$f''''(x) = 120$$

$$f'''''(x) = 0$$

该函数五次以上的导数都为 0。

第 4 节 | 最优化

无论是在统计学还是其他方面，导数的一个重要用途就是求最大化和最小化问题，换句话说，即求函数的最大值和最小值（例如，最大似然法估计、最小二乘法估计）。这些问题统一被称为“最优化”。

如图 2.12 所示，如果函数处于相对（局部）最大值或最小值（即该数值大于或者小于周围的数值），或者处于绝对（全局）最大值或最小值（即该值至少跟其他数值一样大或者一样小），处于该点的切线是水平的，所以函数在该点的导数为 0。

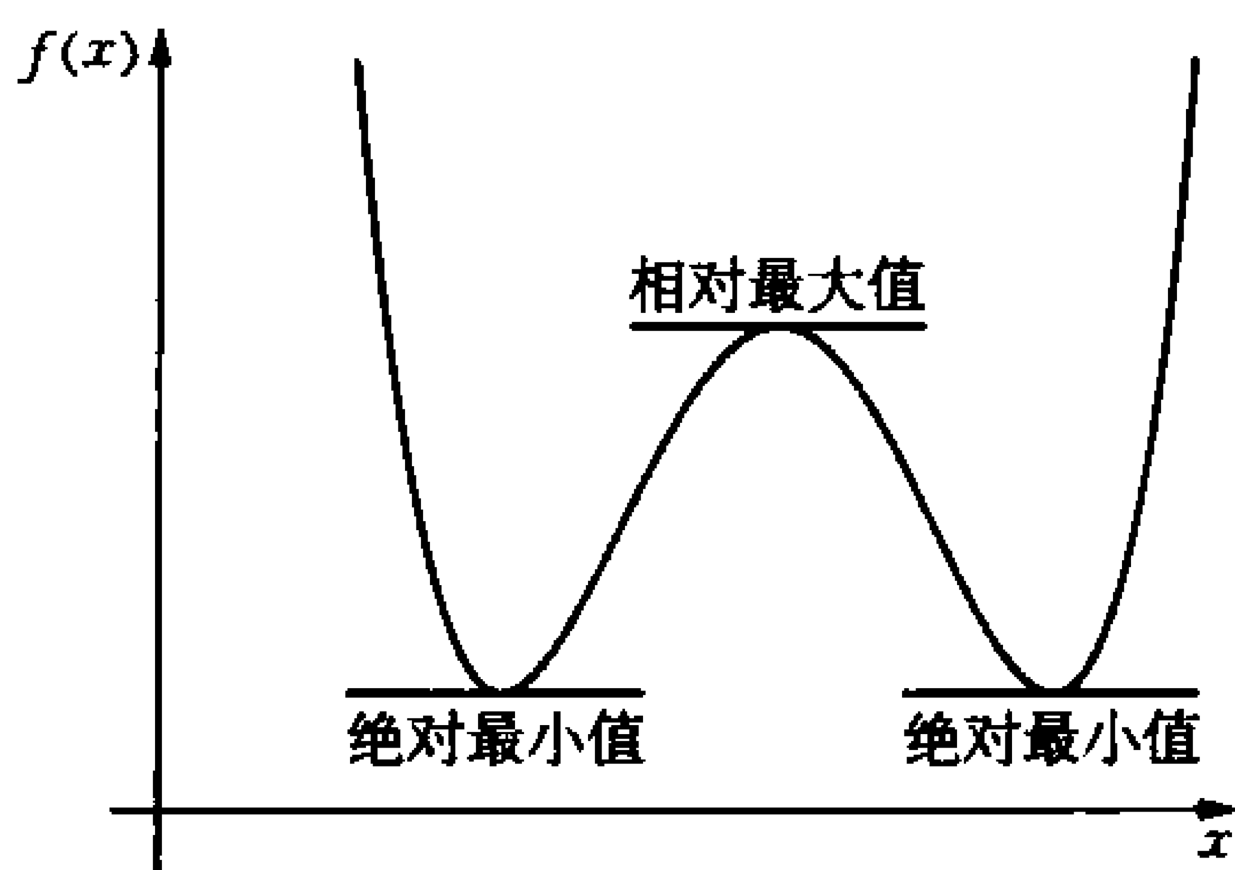


图 2.12　函数的导数为 0 的点是函数 $f(x)$ 的最大值或最小值

但是，导数为 0 的点并不一定是函数的最大值或者最小值。如图 2.13 所示，拐点（函数弯曲方向发生变化的点）的导数同样为 0。导数为 0 的点统称为“驻点”。

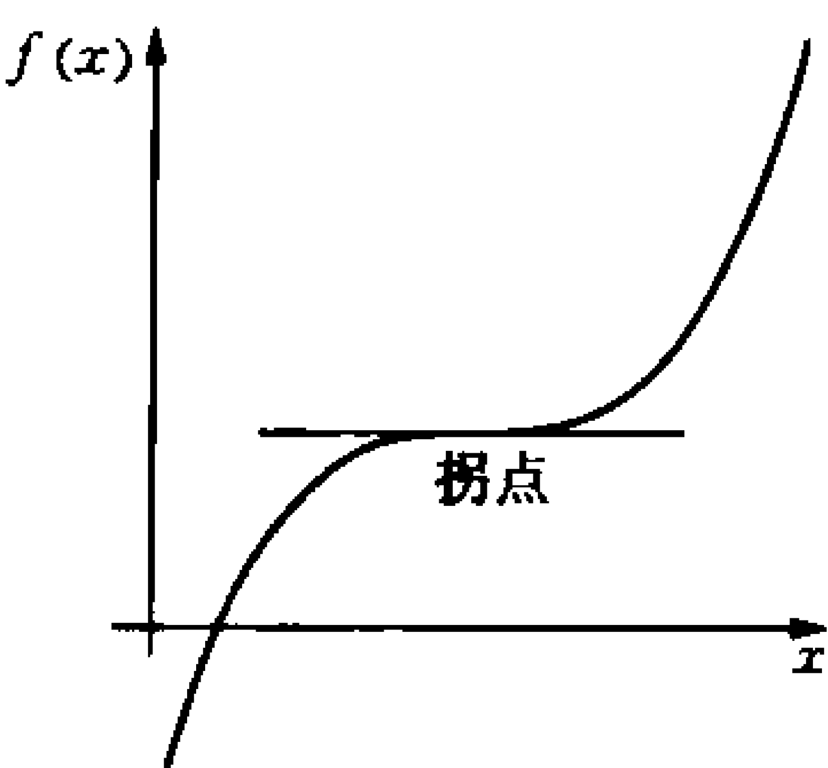
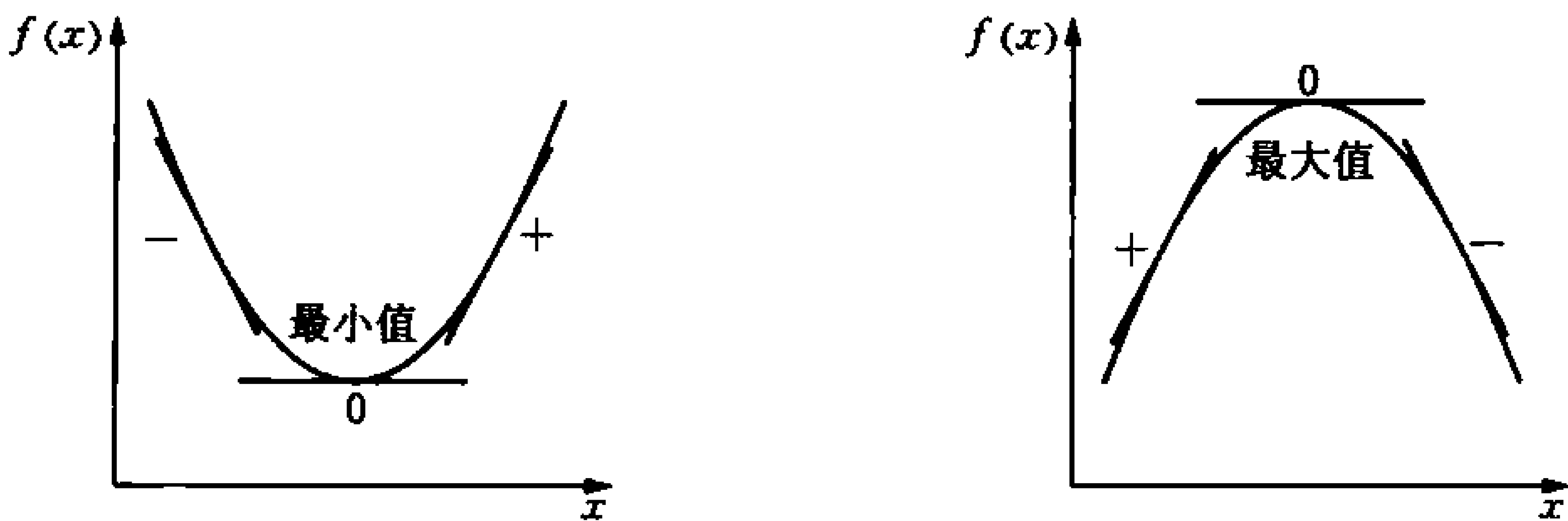


图 2.13 导数为 0 的点是函数 $f(x)$ 的拐点

为了区别导数为 0 的三种情况——最小值、最大值、拐点，我们可以借助二阶导数(如图 2.14)。

原始函数、一阶导数、二阶导数的关系如图 2.15 所示：$f(x)$ 的一阶导数在两个最小值和一个(相对)最大值处为 0$\left(\frac{dy}{dx}=0\right)$；



注：在最小值处，一阶倒数 $f'(x)$ 从负值由 0 变成正值，即一阶导数是递增的，因而二阶导数 $f''(x)$ 是正的。正如一阶导数标示原来函数的变化一样，二阶导数可以标示出一阶导数的变化。在最大值处，一阶倒数 $f'(x)$ 从正值由 0 变成负值，即一阶导数是递减的，因而二阶导数 $f''(x)$ 是负的。而在拐点处，二阶导数 $f''(x)=0$。

图 2.14 若函数 $f(x)$ 在最低点，随着 x 变大，其一阶导数变大；若函数 $f(x)$ 在最高点，随着 x 变大，其一阶导数变小

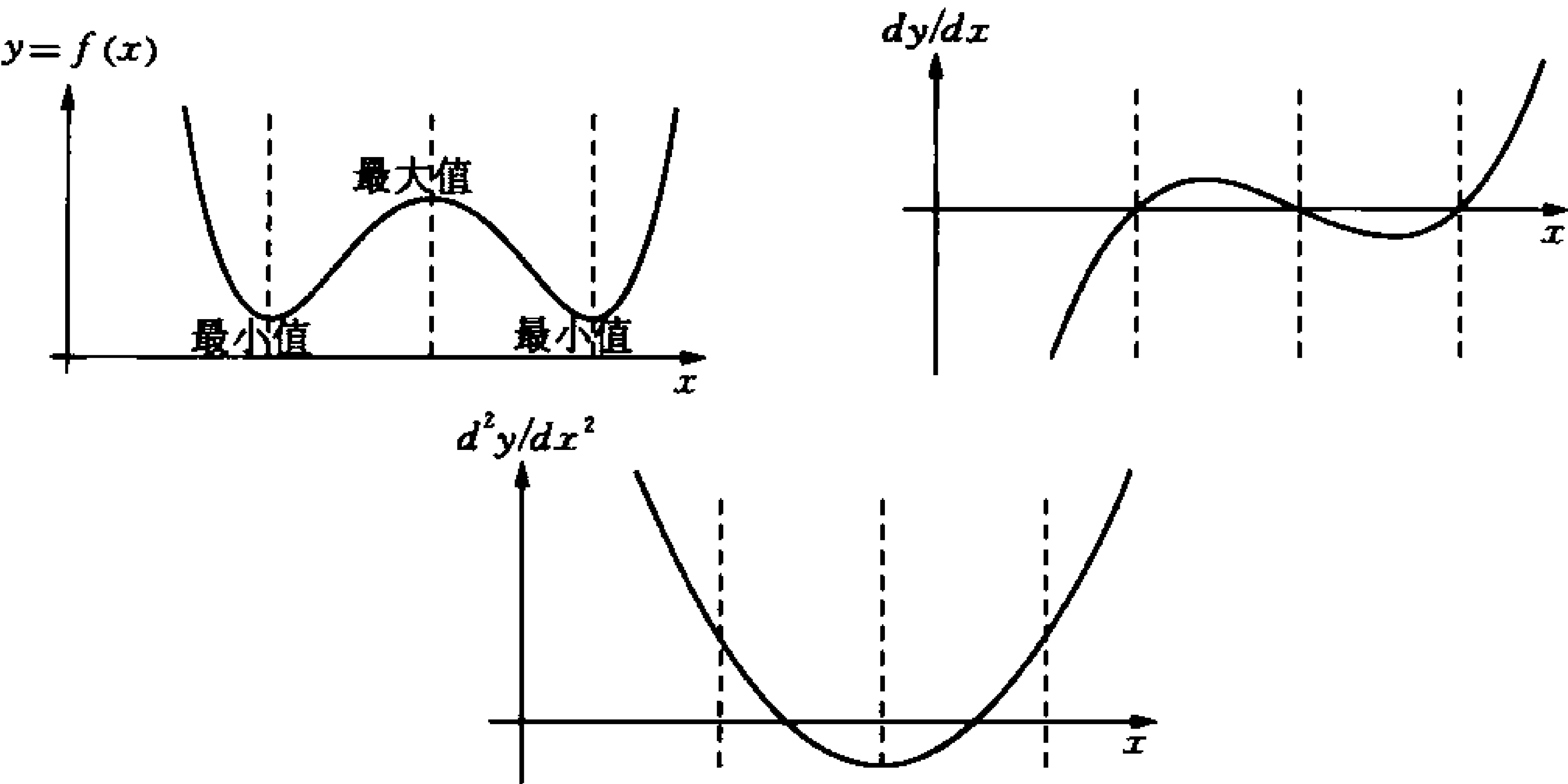


图 2.15 函数的一阶导数和二阶导数

二阶导数 d^2y/dx^2 在两个最小值处为正值，而在最大值处为负值。

最优化的例子

求下面这个函数的极值(最小值或最大值)：

$$f(x) = 2x^3 - 9x^2 + 12x + 6$$

该函数如图 2.16 所示(顺便提一下，确定局部驻点和确定它们是最小值还是最大值，对于函数作图很有用)。

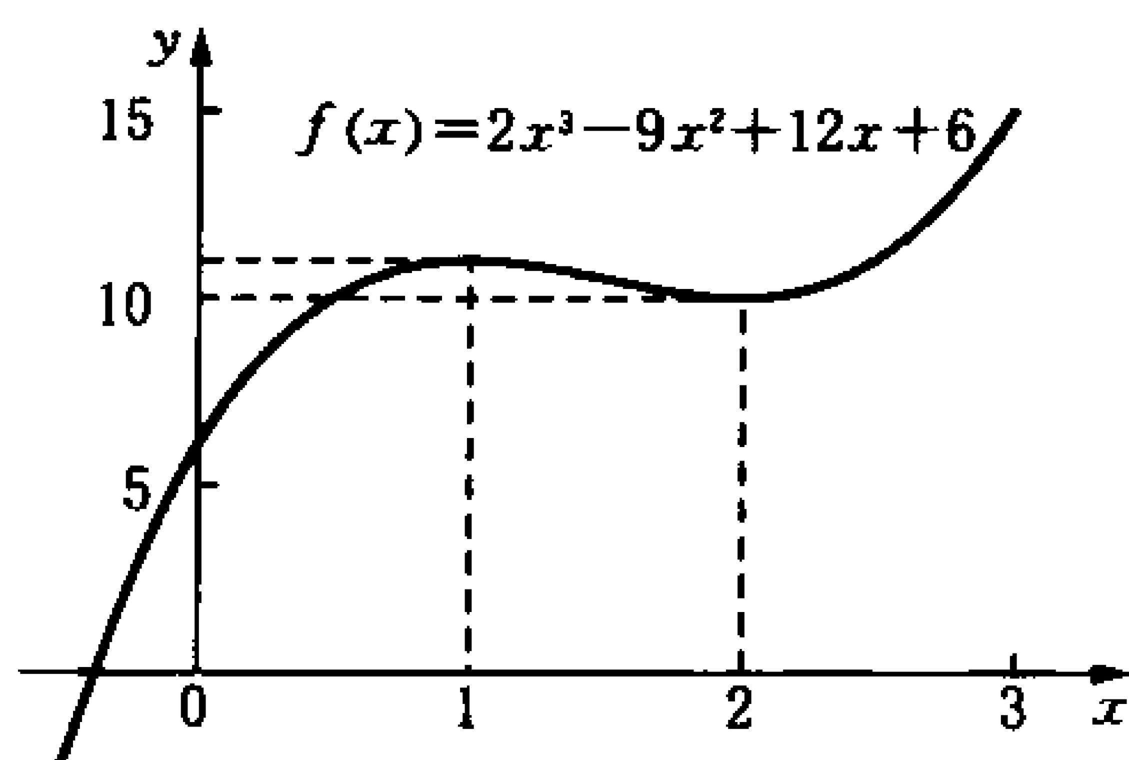


图 2.16 求函数 $f(x) = 2x^3 - 9x^2 + 12x + 6$ 的极值

函数的一阶导数、二阶导数为：

$$f'(x) = 6x^2 - 18x + 12$$

$$f''(x) = 12x^2 - 18$$

令一阶导数等于 0，然后求相应的 x 值得：

$$6x^2 - 18x + 12 = 0$$

$$\Rightarrow x^2 - 3x + 2 = 0$$

$$\Rightarrow (x - 2)(x - 1) = 0$$

$f'(x) = 0$ 的两个根为 $x = 1$ 和 $x = 2$。

对于 $x=2$，

$$f(2)=2\times2^3-9\times2^2+12\times2+6=10$$

$$f'(2)=6\times2^2-18\times2+12=0\checkmark$$

$$f''(2)=12\times2^2-18=6$$

因为 $f''(2)$ 为正值，所以点(2，10)代表了一个(相对)最小值。

对于 $x=1$，

$$f(1)=2\times1^3-9\times1^2+12\times1+6=11$$

$$f'(1)=6\times1^2-18\times1+12=0\checkmark$$

$$f''(1)=12\times1^2-18=-6$$

因为 $f''(1)$ 为负值，所以点(1，11)代表了一个(相对)最大值。

第 5 节 | 多变量和矩阵的微分学

多变量的微分学在统计学中有着广泛的应用。多变量的微分学的关键思想非常直接,即它是单一自变量微分学的扩展,然而该话题在微积分入门介绍中经常被忽略。

偏导数

对于一个具有多个自变量的函数 $y = f(x_1, x_2, \cdots, x_n)$, y 对于 x_i 的偏导数即假定其他 x 为常数时, $f(x_1, x_2, \cdots, x_n)$ 的导数。为了将它和常用导数 dy/dx 相区别,我们常用∂替代 d 来表示偏导数:$\partial y/\partial x_i$。

例如,已知函数

$$y = f(x_1, x_2) = x_1^2 + 3x_1 x_2^2 + x_2^3 + 6$$

该函数对于 x_1 和 x_2 的偏导数为:

$$\frac{\partial y}{\partial x_1} = 2x_1 + 3x_2^2 + 0 + 0 = 2x_1 + 3x_2^2$$

$$\frac{\partial y}{\partial x_2} = 0 + 6x_1 x_2 + 3x_2^2 + 0 = 6x_1 x_2 + 3x_2^2$$

求对于 x_i 的偏导数的“技巧”在于把其他 x 当做常数。所以,当求 y 对于 x_1 的偏导数时 ,x_2^2 和 x_2^3 等均可被当做常数。

偏导数$\partial f(x_1, x_2, \cdots, x_n)/\partial x_1$给出了函数$f(x_1, x_2, \cdots, x_n)$在$x_1$方向上的切超平面。[18]例如，函数$f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 + 10$在$x_1=1$和$x_2=2$时的切面如图 2.17 所示。

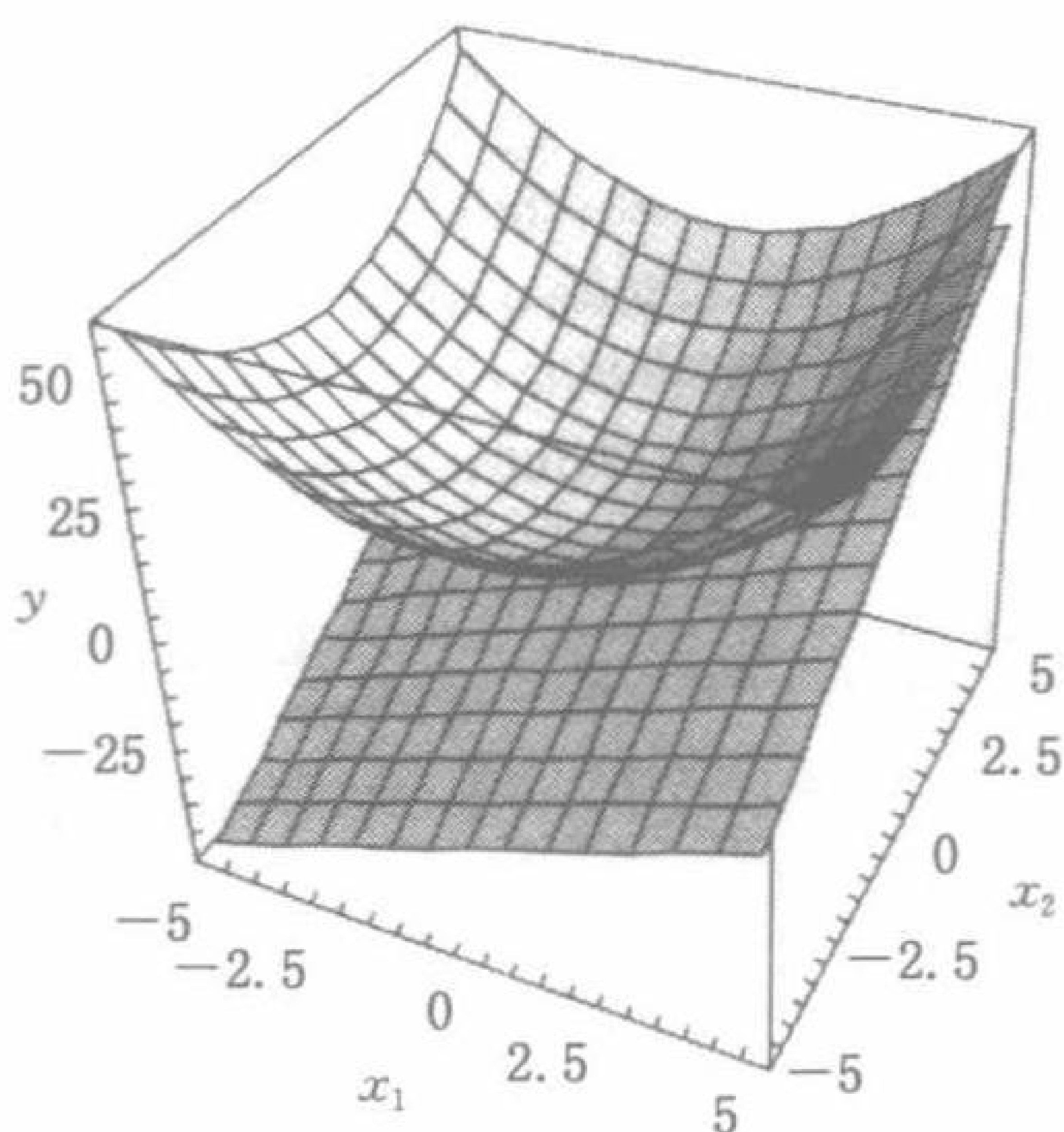


图 2.17　函数$f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 + 10$

在$x_1 = 1$和$x_2 = 2$时的切面

在局部/全局最小值或者最大值中，切面在各个方向上的斜率都为 0。因此，要求一个多变量函数的最小值或最大值，我们就必须分别对每个变量求偏导，使之分别为 0，然后解方程组。

假设我们想寻找x_1和x_2的值，使得函数$f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 + 10$最小。首先我们分别对$x_1$和$x_2$求导：

$$\frac{\partial y}{\partial x_1} = 2x_1 + x_2$$

$$\frac{\partial y}{\partial x_2} = x_1 + 2x_2$$

当偏导数等于 0 时，我们可以得到唯一解：$x_1 = 0$，$x_2 = 0$。在这个例子中，答案相当简单，因为偏导数是x_1、x_2的线性

函数。当函数最小时,其值 $y = 0^2 + 0 \times 0 + 0^2 + 10 = 10$。

如图2.17所示,在 $x_1 = 1$ 和 $x_2 = 2$ 以上的切面斜率为:

$$\frac{\partial y}{\partial x_1} = 2(1) + 2 = 4$$

$$\frac{\partial y}{\partial x_2} = 1 + 2(2) = 5$$

拉格朗日乘数和受约束的最优化

拉格朗日乘数使我们能在条件 $g(x_1, x_2, \cdots, x_n) = 0$ 下最优化函数 $y = f(x_1, x_2, \cdots, x_n)$。这种方法实际上是在偏导数中加入了限制。

举个简单的例子:将函数 $y = f(x_1, x_2) = x_1^2 + x_2^2$ 最小化是要受条件 $x_1 + x_2 = 1$ 制约的(假如没有该约束条件,显然 $x_1 = x_2 = 0$ 时函数最小)。解决受约束的最小化问题的方法如下:

第一,将约束条件移项成标准形式 $g(x_1, x_2, \cdots, x_n) = 0$,得 $x_1 + x_2 - 1 = 0$。

第二,构造一个具有如下标准形式的新方程[19]:

$$h(x_1, x_2, \cdots, x_n, \lambda) \equiv f(x_1, x_2, \cdots, x_n) - \lambda \times g(x_1, x_2, \cdots, x_n)$$

新的自变量 λ 读作"拉格朗日乘数",在这个例子中,

$$h(x_1, x_2, \lambda) \equiv x_1^2 + x_2^2 - \lambda(x_1 + x_2 - 1)$$

第三,寻找最优化函数 $h(x_1, x_2, \cdots, x_n, \lambda)$ 的 x_1, x_2, …, x_n, λ 值,即让 $h(x_1, x_2, \cdots, x_n, \lambda)$ 分别对 x_1,

x_2, …, x_n, λ 求偏导，把这个 $n+1$ 个偏导数都设为 0，然后解方程组求得 x_1, x_2, …, x_n, λ。在此例中，

$$\frac{\partial h(x_1, x_2, \lambda)}{\partial x_1} \equiv 2x_1 - \lambda$$

$$\frac{\partial h(x_1, x_2, \lambda)}{\partial x_2} \equiv 2x_2 - \lambda$$

$$\frac{\partial h(x_1, x_2, \lambda)}{\partial \lambda} \equiv - x_1 - x_2 + 1$$

注意，令 λ 偏导为 0 所得到的等式即约束条件 $x_1 + x_2 - 1 = 0$。因此，所有满足偏导为 0 的解必须首先满足约束条件。所以在本例中，其存在唯一解：

$$x_1 = x_2 = 0.5 \quad (\lambda = 1)$$

拉格朗日乘数可以解决多个约束条件的问题，只要我们给每个约束条件引入一个拉格朗日乘数即可。

矩阵的微分

对于自变量为 x_1, x_2, …, x_n 的函数 $y = f(x_1, x_2, \cdots, x_n)$，我们可以将其简化为 $y = f(\mathbf{x})$。其中，向量 $\mathbf{x} = [x_1, x_2, \cdots, x_n]'$。$y$ 关于 $\mathbf{x}$ 的向量偏导数（或者梯度）是指 y 对于每一个列向量元的偏导数。

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \dfrac{\partial y}{\partial x_1} \\ \dfrac{\partial y}{\partial x_2} \\ \vdots \\ \dfrac{\partial y}{\partial x_n} \end{bmatrix}$$

如果 y 是 $\mathbf{x}$ 的线性方程，

$$y = \underset{(1\times n)}{\mathbf{a}'}\underset{(n\times 1)}{\mathbf{x}} = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

那么，$\partial y/\partial x_i = a_i$，$\partial y/\partial \mathbf{x} = \mathbf{a}$，例如，

$$y = x_1 + 3x_2 - 5x_3 = [1,\ 3,\ -5]\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

向量的偏导数为：

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} 1 \\ 3 \\ -5 \end{bmatrix}$$

如果 y 是 $\mathbf{x}$ 的二次形式，则有：

$$y = \underset{(1\times n)}{\mathbf{x}'}\underset{(n\times n)}{\mathbf{A}}\underset{(n\times 1)}{\mathbf{x}}$$

其中，矩阵 $\mathbf{A}$ 是一个对称矩阵。把矩阵乘积展开后，得到：

$$y = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + \cdots$$
$$+ 2a_{1n}x_1x_n + \cdots + 2a_{n-1,\,n}x_{n-1}x_n$$

因此，

$$\frac{\partial y}{\partial x_i} = 2(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n) = 2\mathbf{a}_i'\mathbf{x}$$

其中，$\mathbf{a}_i'$代表 $\mathbf{A}$ 的第 i 行。把这些偏导数写成向量形式，即 $\partial y/\partial \mathbf{x} = 2\mathbf{A}\mathbf{x}$。线性函数和二次函数的向量偏导数与单变量函数的纯量偏导数是一样的：$d(ax)/dx = a$，$d(ax^2)/dx = 2ax$。

例如，对于

$$y = [x_1,\ x_2]\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 2x_1^2 + 3x_1x_2 + 3x_2x_1 + x_2^2$$

$$= 2x_1^2 + 6x_1x_2 + x_2^2$$

其对 x_1 和 x_2 的偏导数为：

$$\frac{\partial y}{\partial x_1} = 4x_1 + 6x_2$$

$$\frac{\partial y}{\partial x_2} = 6x_1 + 2x_2$$

那么，向量的偏导数为：

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} 4x_1 + 6x_2 \\ 6x_1 + 2x_2 \end{bmatrix} = 2\begin{bmatrix} 2 & 3 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \checkmark$$

$y = f(\mathbf{x})$ 的二阶偏导数——海森矩阵的定义如下：

$$\frac{\partial^2 y}{\partial \mathbf{x} \partial \mathbf{x}'} = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \frac{\partial^2 y}{\partial x_2 \partial x_1} & \frac{\partial^2 y}{\partial x_2^2} & \cdots & \frac{\partial^2 y}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \frac{\partial^2 y}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 y}{\partial x_n^2} \end{bmatrix}$$

对于对称矩阵 $\mathbf{A}$，$\partial^2(\mathbf{x}'\mathbf{A}\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}' = 2\mathbf{A}$。

为了使多变量函数 $y = f(\mathbf{x})$ 最小，我们可以将向量偏导数设为 $\mathbf{0}$，即 $\partial y/\partial \mathbf{x} = \mathbf{0}$，然后解相应的关于 $\mathbf{x}$ 的方程组，得到解 $\mathbf{x}^*$。如果海森矩阵在 $\mathbf{x} = \mathbf{x}^*$ 时是正定的，那么该解代表函数的一个(局部)最小值；如果海森矩阵是负定的，那么该解代表函数的一个最大值。[20]这与单变量函数的导数相同，即二阶导数 d^2y/dx^2 为最小值时是正的，为最大值时是

负的。

如之前的函数，

$$y = f(x_1, x_2) = x_1^2 + x_1x_2 + x_2^2 + 10$$

在 $x_1 = x_2 = 0.5$ 处有一个驻点(即在该点上，其偏导数为0)，那么，函数的二阶偏导数为：

$$\frac{\partial^2 y}{\partial x_1 \partial x_2} = \frac{\partial^2 y}{\partial x_2 \partial x_1} = 1$$

$$\frac{\partial^2 y}{\partial x_1^2} = \frac{\partial^2 y}{\partial x_2^2} = 2$$

因此，在 $x_1 = x_2 = 0.5$ (或者其他点)时的海森矩阵如下：

$$\begin{bmatrix} \frac{\partial^2 y}{\partial x_1^2} & \frac{\partial^2 y}{\partial x_1 \partial x_2} \\ \frac{\partial^2 y}{\partial x_2 \partial x_1} & \frac{\partial^2 y}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

明显是正定的，那么，我们可以证明在 $x_1 = x_2 = 0.5$ 时，$y = 10$ 是 $f(x_1, x_2)$ 的一个最小值。

第6节｜泰勒展式

假如一个函数 $f(x)$ 在 $x = x_0$ 处拥有无限阶导数(尽管大部分可能是0)，那么该函数可以分解成泰勒展式：

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots$$

$$= \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \qquad [2.1]$$

其中，$f^{(n)}$ 表示 f 的 n 阶导数，$n!$ 表示 n 的阶乘。[21]

只要 x 充分接近 x_0，同时函数 $f(\cdot)$ 足够大，那么，我们只需要取泰勒展式的前几项就可能接近 $f(x)$。例如，函数 $f(x)$ 在 x 与 x_0 之间是二次型的，那么，$f(x)$ 就可以近似等于泰勒展式的前三项，剩下的导数会很小，可以忽略不计。同样，如果函数 $f(x)$ 在 x 和 x_0 之间是线性的，那么，$f(x)$ 就可以近似为泰勒展式的前两项。

我们可以通过下面的三次函数来了解泰勒展式的应用：

$$f(x) = 1 + x^2 + x^3$$

那么，我们有：

$$f'(x) = 2x + 3x^2$$

$$f''(x) = 2 + 6x$$

$$f'''(x) = 6$$

$$f^{(n)}(x) = 0 \quad (n > 3)$$

取 $x_0 = 2$，求得各阶导数的值分别为：

$$f(2) = 1 + (2)^2 + (2)^3 = 13$$

$$f'(2) = 2(2) + 3(2)^2 = 16$$

$$f''(2) = 2 + 6(2) = 14$$

$$f'''(2) = 6$$

最后，我们利用 $x_0 = 2$ 时的泰勒展式来求 $f(x)$ 在 $x = 4$ 时的值：

$$f(4) = f(2) + \frac{f'(2)}{1!}(4-2) + \frac{f''(2)}{2!}(4-2)^2 + \frac{f''(2)}{3!}(4-2)^3$$

$$= 13 + 16(2) + \frac{14}{2}(2^2) + \frac{6}{6}(2^3)$$

$$= 81$$

将 $x = 4$ 代入原函数直接检验得：

$$f(4) = 1 + 4^2 + 4^3 = 81$$

在这个例子中，如果取少于 4 项的泰勒展式，就会得到一个很差的近似（因为这是一个三次函数）。

泰勒展式和近似可以扩展到多变量函数中，当函数是纯量函数或者我们可以应用一阶近似或二阶近似时，问题就会变得很简单。假设 $y = f(x_1, x_2, \cdots, x_n) = f(\mathbf{x})$，同时我们想知道 $f(\mathbf{x})$在 $\mathbf{x} = \mathbf{x}_0$ 处的近似，那么$f(\mathbf{x})$的二阶泰勒展式可近似为：

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + [g(\mathbf{x}_0)]'(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)'\mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

其中，$f(\mathbf{x})$ 的梯度 $g(\mathbf{x}) \equiv \partial y/\partial \mathbf{x}$，海森矩阵 $\mathbf{H}(\mathbf{x}) \equiv \partial^2 y/\partial \mathbf{x}\partial \mathbf{x}'$，它们都是在 $\mathbf{x} = \mathbf{x}_0$ 的情况下估计的。我们可以发现，这个展式和方程 2.1 给出的纯量泰勒展式非常相似。

第 7 节 | 积分学的基本思想

面积:定积分

如图 2.18,我们首先考虑一下曲线 $f(x)$下水平坐标 x_0 和 x_1 间所包含的面积。这个面积可以由以下近似求得:把 x_0 和 x_1 之间的线段分成 n 等分,每段长度为 Δx,并分别和曲线 $f(x)$连接,构造成一系列长方形,如图 2.19 所示。那么,各个长方形底边所对应的 x 坐标分别为:

$$x_0,\ x_0+\Delta x,\ x_0+2\Delta x,\ \cdots,\ x_0+n\Delta x$$

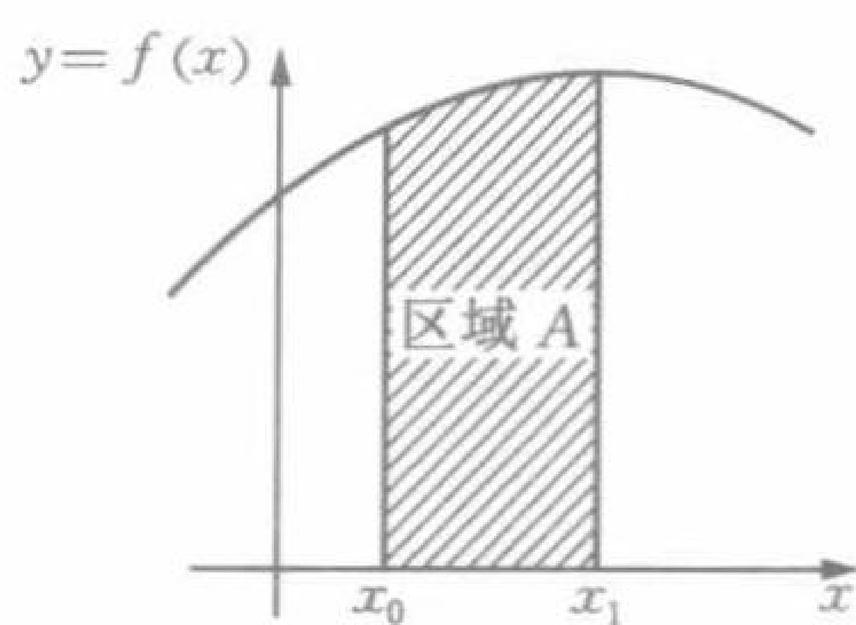


图 2.18　函数 $f(x)$在 x_0 和 x_1 之间的区域

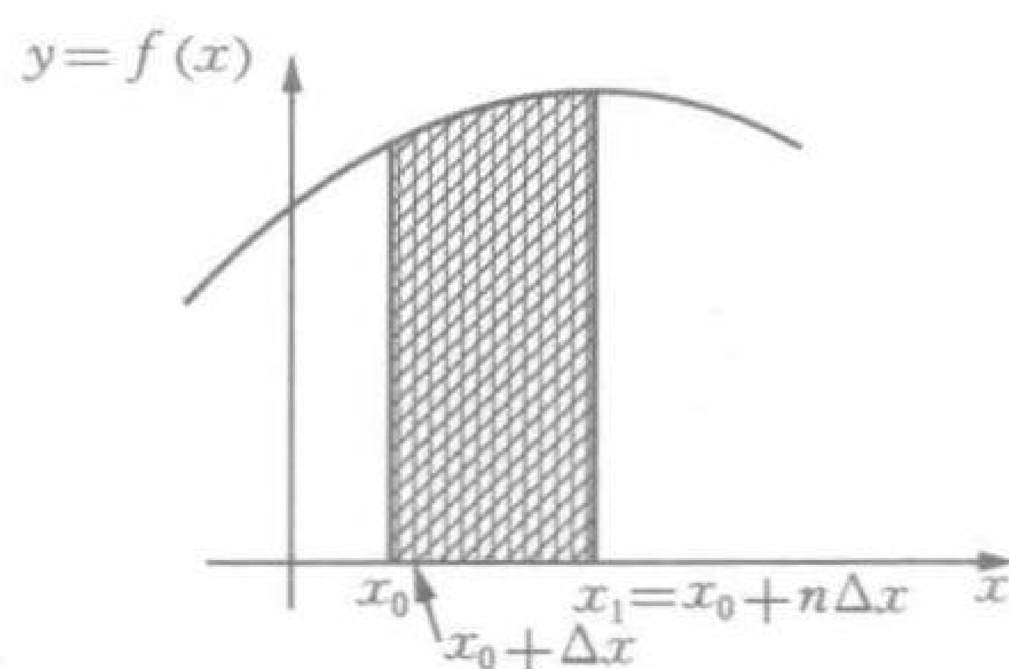


图 2.19　曲线以下区域可看做无数长方形区域之和

因此,所有长方形面积之和为:

$$\sum_{i=0}^{n-1} f(x_0+i\Delta x)\Delta x \approx A$$

且面积的近似值会随着 n 值的增大而越来越精确。用极限表示为[22]：

$$A=\lim_{\substack{\Delta x\to 0\\ n\to\infty}}\sum_{i=0}^{n-1}f(x_0+i\Delta x)\Delta x$$

该极限可以表示为 $A=\int_{x_0}^{x_1}f(x)dx$，读作“$f(x)$在 $x=x_0$ 到 $x=x_1$ 的定积分”。在这里，x_0、x_1 是积分域，dx 是长方形长度 Δx 无限小的量。积分符号 $\int$ 是拉长的“S”，其所表示的定积分可以理解为连续求和。

如图 2.20 所示，定积分同时确定了面积的符号，如果 y 包含一些小于 0 的值，那么面积可能为负。

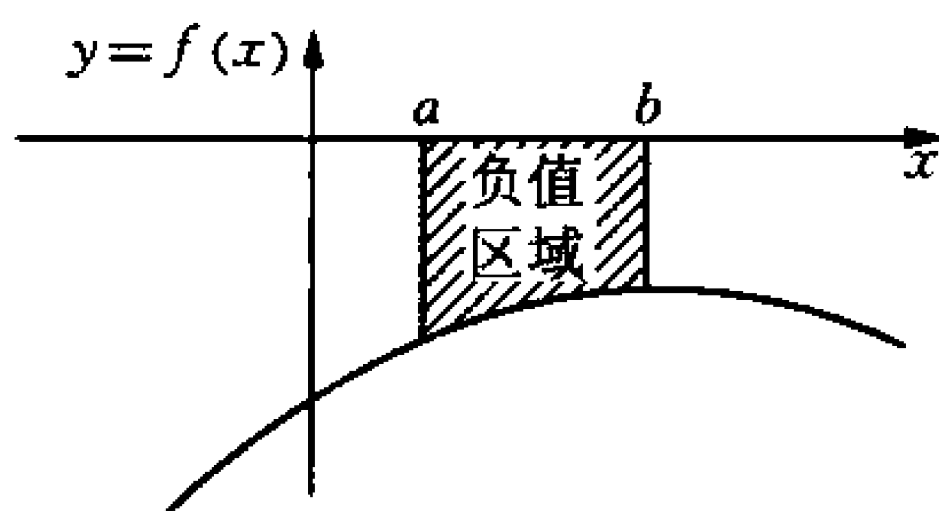


图 2.20　积分 $\int_a^b f(x)dx$ 为负（因为其 y 值在上下限 a 和 b 之间为负）

不定积分

假设对于函数 $f(x)$，存在另一个函数 $F(x)$，使得：

$$\frac{dF(x)}{dx}=f(x)$$

即 $f(x)$是 $F(x)$的导数，那么，$F(x)$就叫做“$f(x)$的反导数”或者“不定积分”。

一个函数的不定积分不是唯一的，因为假如 $F(x)$是

$f(x)$的反导数,那么,$G(x)=F(x)+c$也是(其中,c是绝对常数而不是x的函数)。相反,假如$F(x)$和$G(x)$均为$f(x)$的反导数,那么则存在常数c,使得$G(x)=F(x)+c$。

例如,$f(x)=x^3$,函数$\frac{1}{4}x^4+10$是$f(x)$的反导数,$\frac{1}{4}x^4-10$和$\frac{1}{4}x^4$也是其反导数。事实上,任何$F(x)=\frac{1}{4}x^4+c$形式的函数都是其反导数。

对于不定积分,我们可以写出:

$$\frac{dF(x)}{dx}=f(x)$$

$$F(x)=\int f(x)dx$$

积分符号在定积分和不定积分中的应用是一致的,并且都称为"积分"(这将在下文叙述)。但是,在不定积分中,积分符号上没有积分域,同时请注意,定积分所包含的面积是一个特定的数字,而不定积分是一个函数。

微积分的基本定理

牛顿和莱布尼茨指出,曲线的反导数和曲线以下的面积存在一系列的关系。我们把他们所发现的这种不定积分和定积分之间的关系称为"微积分基本定理":

$$\int_{x_0}^{x_1} f(x)dx=F(x_1)-F(x_0)$$

其中,$F(\cdot)$是$f(\cdot)$的反导数。

以下是一个关于该定理的不严格证明:如图2.21所示,

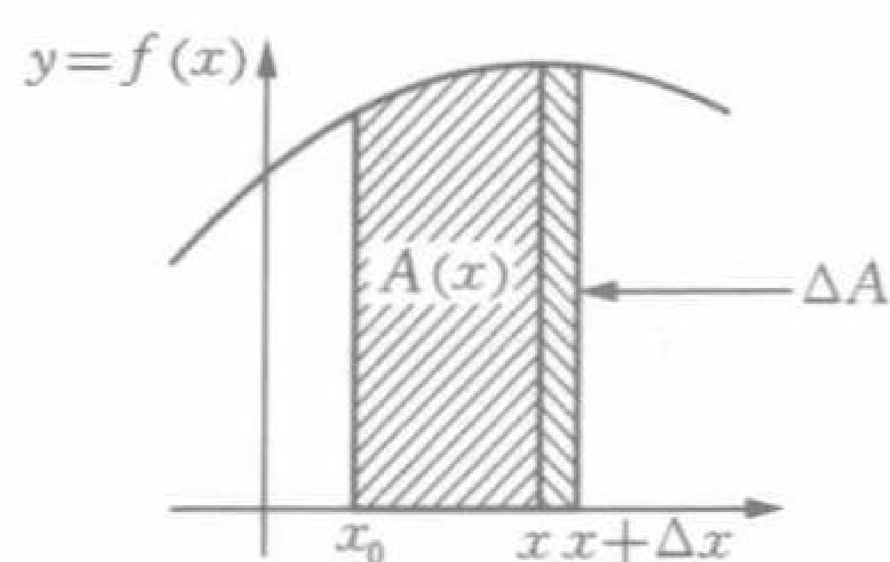


图 2.21 函数 $A(x)$ 以下 x_0 到 x 之间的区域

考虑曲线 $f(x)$ 下一个定点 x_0 和一个动点 x 之间的面积 $A(x)$。$A(x)$ 表明面积是 x 的函数：面积随着 x 由左移向右而改变。在图 2.21 中，$x+\Delta x$ 表示一个比 x 稍微偏右的值，ΔA 表示 x 和 $x+\Delta x$ 之间的面积，这个面积可以近似地看做一个长方形的面积：

$$\Delta A \approx f(x)\Delta x$$

同时，我们可以把该面积表示为：

$$\Delta A = A(x+\Delta x) - A(x)$$

求 A 关于 x 的导数，得到：

$$\frac{dA(x)}{dx} = \lim_{\Delta x \to 0} \frac{\Delta A}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x)\Delta x}{\Delta x} = f(x)$$

最后，$A(x) = \int f(x)dx$ 是 $f(x)$ 的一个特别但未知的不定积分，$F(x)$ 是 $f(x)$ 其他特定的、人为给定的不定积分。对于一些 c 值，$A(x) = F(x) + c$（如前所述，同一函数的两个不定积分的区别在于常数 c）。我们知道，$A(x_0) = 0$，其原因在于，$A(x)$ 表示曲线在 x_0 和 x 之间的面积，而 x_0 和 x_0 之间的面积为 0，所以，

$$A(x_0) = F(x_0) + c = 0$$

$$\Rightarrow c = -F(x_0)$$

因此，对于特定值 $x = x_1$，

$$A(x_1) = \int_{x_0}^{x_1} f(x)dx = F(x_1) - F(x_0)$$

其中，$F(\cdot)$是 $f(\cdot)$的反导数。

例如，我们想知道面积(定积分) $A=\int_1^3(x^2+3)dx$，该面积如图 2.22 所示，我们可以方便地选择 $F(x)=\frac{1}{3}x^3+3x$。[23]

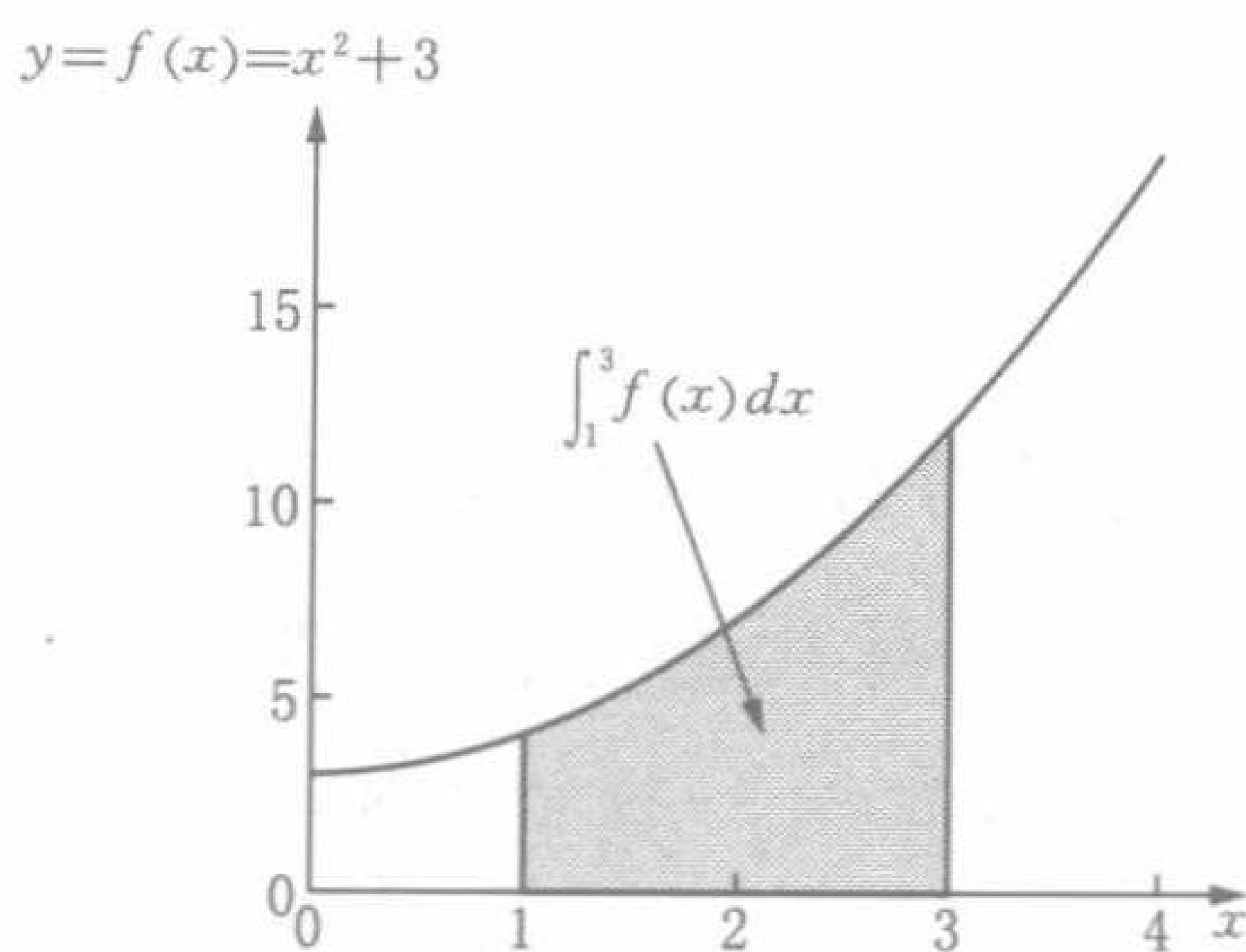


图 2.22　$A=\int_1^3(x^2+3)dx$ 所代表的区域

那么，

$$
\begin{aligned}
A &= F(3)-F(1) \\
&= \left(\frac{1}{3}3^3+3\times 3\right)-\left(\frac{1}{3}1^3+3\times 1\right) \\
&= 18-3\frac{1}{3}=14\frac{2}{3}
\end{aligned}
$$

第8节 | 推荐阅读

关于微积分入门的书目种类繁多，而我仅仅读过其中的一小部分。当然，我最喜欢的是汤普森(Thompson)和加德纳(Garnder)(1998)的著作。关于多变量微积分学在社会科学应用中的进一步的讨论，可参见宾默尔(Binmore)和戴维斯(Davies)(2001)的研究。

第3章

概率估计

本章对应用统计学中广泛运用的概率及统计推理进行了概述。我们知道，初等统计课程，尤其是社会学专业开设的初等统计课程，对概率估计理论仅仅提供了简单的框架介绍。然而，深入了解并熟悉有关话题的背景知识是相当必要的。

第1节 | 初等概率理论

概率基础

在概率理论中,实验是对观察的可重复验证过程;结果是通过对一个可能的观察进行实验所得出的结论;实验的样本空间则为所有可能结果的集合。实验的任何特定“实现”都会在样本空间中产生一个特定的结果。样本空间可以是离散且有限的,或者是离散且无限的,也可以是连续的。例如,掷两次硬币,然后记录下每次投掷的结果(出现的是硬币的正面还是反面)。对于此例,实验的样本空间是离散且有限的,其结果组合为 $S=\{HH, HT, TH, TT\}$。如果我们反复掷硬币,并记录每次投掷的结果,此时样本空间是离散且无限的,其包括的正整数组合有 $S=\{1, 2, 3, \cdots\}$。[24]如果我们把灯泡一直开着直到保险丝烧断,并记录下灯泡从打开一直到自然熄灭所需要的时间,此时实验的样本空间就是连续的,其包括所有的正实数(这里无需明确指出灯泡寿命的上限):$S=\{x:x>0\}$。在本节中,我叙述的内容仅限于样本空间是离散且有限的情况。

一个事件是实验的样本空间子集,即结果集合。如果包含在结果集合中的情况发生,我们就说该事件在实验中发

生。例如，对于 $S=\{HH, HT, TH, TT\}$，如果我们得到结果 HH 或 HT，则事件 $E\equiv\{HH, HT\}$（代表第一次掷硬币出现正面）发生。请注意，通过以上定义，样本空间 S 本身和不包含任何事件的零或空事件 $\phi\equiv\{\}$ 都是事件。

概率论定理

令 $S=\{o_1, o_2, \cdots, o_n\}$ 表示实验的样本空间；$O_1=\{o_1\}$，$O_2=\{o_2\}$，…，$O_n=\{o_n\}$ 表示单一事件，且每个事件包含一个结果；事件 $E=\{o_a, o_b, \cdots, o_m\}$，为 S 的一个子空间（下标 $a, b, \cdots, m$ 是 1 到 n 之间的不同数字）。概率是满足如下定理的事件所发生的可能性[25]：

P1：$\Pr(E)\geqslant 0$：一个事件发生的概率是非负的；

P2：$\Pr(E)=\Pr(O_a)+\Pr(O_b)+\cdots+\Pr(O_m)$，一个事件发生的概率为所有构成其结果的和。

P3：$\Pr(S)=1$ 和 $\Pr(\varnothing)=0$：样本空间是穷尽的，即某些事件必然发生。

假设样本空间 $S=\{HH, HT, TH, TT\}$ 包含所有结果，且每个结果发生的可能性相同，即：

$$\Pr(HH)=\Pr(HT)=\Pr(TH)=\Pr(TT)=0.25$$

那么，对于事件 $E=\{HH, HT\}$，$\Pr(E)=0.25+0.25=0.5$。这个例子比较简单，因为每个结果发生的概率都相同，正如扔硬币得到正反面的概率相同一样。实际上，只要各结果发生概率之和为 1，即符合以上定理。

在经典统计学中，并且从大多数统计学应用的角度来看，概率是指长期的均衡比例。即，假如一个事件发生的概率为 $\frac{1}{2}$，那么当实验重复多次，这个事件发生的概率会接近于0.5，且这个接近过程会随重复次数的增加而越发完善。这是客观论者对概率的一般性理解：概率为长期的相对频率，即均衡比率。

事件之间的关系、条件概率与独立事件

事件之间存在许多重要的关系。两个事件 E_1 和 E_2 的交集，记做 $E_1 \cap E_2$，它包括两个事件中共有的所有结果。因此，$\Pr(E_1 \cap E_2)$ 表示 E_1 和 E_2 同时发生的概率。如果 $E_1 \cap E_2 = \varnothing$，则称 E_1 和 E_2 无交集或者互斥。推广后，可以知道，一系列事件的交集 $E_1 \cap E_2 \cap \cdots \cap E_k$ 包含事件 E_1 到事件 E_k 共有的所有结果。例如，我们有事件 $E_1 \equiv \{HH, HT\}$（第一次掷硬币出现正面）、$E_2 \equiv \{HH, TH\}$（第二次掷硬币出现正面）和 $E_3 \equiv \{TH, TT\}$（第一次掷硬币出现反面），那么，可知 $E_1 \cap E_2 = \{HH\}$，$E_1 \cap E_3 = \varnothing$，$E_2 \cap E_3 = \{TH\}$。

两个事件 E_1 和 E_2 的并集 $E_1 \cup E_2$ 包含两个事件中所有的结果；$\Pr(E_1 \cup E_2)$是事件 E_1 或者事件 E_2 发生的概率。那么，事件 $E_1 \cup E_{21} \cup \cdots \cup E_k$ 的并集是 E_1 到 E_k 中含有的所有结果。如果这些事件无交集，那么，

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_k) = \sum_{i=1}^{k} \Pr(E_i)$$

否则，

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_k) < \sum_{i=1}^{k} \Pr(E_i)$$

由于不同事件中所包含的结果可能有重复，因此，某些事件发生的概率之和可能大于1。因此，任意两个事件发生的概率为：

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

即两个事件分别发生的概率之和减去两个事件交集发生的概率（因为在算两个事件分别发生的概率之和时，交集部分被算了两次）。由此，可引申到事件发生概率相同的例子（如前所述，E_1 和 E_2 无交集，而 E_1 和 E_2 有交集）：

$$\begin{aligned}\Pr(E_1 \cup E_3) &= \Pr(HH, HT, TH, TT) = 1\\ &= \Pr(E_1) + \Pr(E_3)\\ &= 0.5 + 0.5\end{aligned}$$

$$\begin{aligned}\Pr(E_1 \cup E_2) &= \Pr(HH, HT, TH) = 0.75\\ &= \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)\\ &= 0.5 + 0.5 - 0.25\end{aligned}$$

给定事件 E_1、事件 E_2 发生的条件概率为：

$$\Pr(E_2 \mid E_1) \equiv \frac{\Pr(E_2 \cap E_1)}{\Pr(E_1)} \qquad [3.1]$$

条件概率可以这样解释：如果已知事件 E_1 会发生，那么，求事件 E_2 发生的概率。为求得 $\Pr(E_2 \cap E_1)$，我们解方程 3.1 即可得到一般概率的乘法法则：

$$\Pr(E_2 \cap E_1) = \Pr(E_1)\Pr(E_2 \mid E_1)$$

交换 E_1 和 E_2 的角色后，得到以下方程：

$$\Pr(E_1 \mid E_2) \equiv \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)} \qquad [3.2]$$

$$\Pr(E_1 \cap E_2) = \Pr(E_2)\Pr(E_1 \mid E_2) \qquad [3.3]$$

如果 $\Pr(E_2 \cap E_1) = \Pr(E_1)\Pr(E_2)$，我们说这两个事件为独立事件。方程 $\Pr(E_2 \cap E_1) = \Pr(E_1)\Pr(E_2)$ 称为“独立事件概率的乘法法则”。事件 E_1 和事件 E_2 的独立性暗示了 $\Pr(E_1) = \Pr(E_1 \mid E_2)$ 和 $\Pr(E_2) = \Pr(E_2 \mid E_1)$，即，两个独立事件的无条件概率与已知其中一个事件会发生时，另一个事件的发生概率相同。推广后可知，若已知一系列独立事件 $\{E_1, E_2, \cdots, E_k\}$，那么，对于发生其中任意两个或多个事件的子集的概率为：

$$\Pr(E_a \cap E_b \cap \cdots \cap E_m) = \Pr(E_a)\Pr(E_b)\cdots\Pr(E_m)$$

因此，若已知第一次掷硬币得到硬币正面，则第二次掷硬币也为正面的概率为：

$$\begin{aligned}\Pr(E_2 \mid E_1) &= \frac{\Pr(E_2 \cap E_1)}{\Pr(E_1)} \\ &= \frac{0.25}{0.5} \\ &= \Pr(E_2)\end{aligned}$$

同理，$\Pr(E_1 \cap E_2) = 0.25 = \Pr(E_1)\Pr(E_2) = 0.5 \times 0.5$。因此，事件 E_1 和事件 E_2 是独立事件。

两个事件独立与两个事件互斥不同，因为两个事件互斥暗示了它们不可能一起发生，所以，它们是互相依赖的。在我们的例子中，事件 E_1 和事件 E_2 是独立但不互斥的：$E_1 \cap E_2 = \{HH\} \neq \varnothing$。

事件 E_1 和事件 E_2 的差包含了所有在事件 E_1 中发生而没有在事件 E_2 中发生的结果，记做 $E_1 - E_2$。那么实验样本空间包含的所有事件与事件 E 的差称为“事件 E 的补集”，且 $\Pr(\bar{E}) = 1 - \Pr(E)$。对于之前提到的例子，结果发生概率相同的事件 $E_1 = \{HH, HT\}$，其补集发生的概率为 $\Pr(\bar{E}_1) = \Pr(TH, TT) = 0.5 = 1 - 0.5$。

Bonferroni 不等式

令 $E \equiv E_1 \cap E_2 \cap \cdots \cap E_k$，那么 $\bar{E} = \bar{E} \cup \bar{E}_2 \cup \cdots \cup \bar{E}_k$，运用之前的方程，则：

$$\Pr(E_1 \cap E_2 \cap \cdots \cap E_k) = \Pr(E) = 1 - \Pr(\bar{E}) \quad [3.4]$$
$$\geqslant 1 - \sum_{i=1}^{k} \Pr(\bar{E}_i)$$

假设所有事件 $E_1, E_2, \cdots, E_k$ 发生的概率都相等，那么，对于任意 E_i，其发生的概率都等于 $\Pr(E_i) = 1 - b$。那么，

$$\Pr(E_1 \cap E_2 \cap \cdots E_k) \equiv 1 - a \quad [3.5]$$
$$\geqslant 1 - kb$$

方程 3.5 与一般方程 3.4 都称为“Boneferroni 不等式”。

方程 3.5 对线性联立方程的应用有以下暗示：假设 b 是每 k 个非独立统计检验的 I 类错误比率（例如，显著水平 a），a 表示合并的 I 类错误比率，即 k 个非独立统计检验中至少错误地拒绝了一个为真的零假设的概率，那么，$a \leqslant kb$。例如，我们在 0.01 显著水平下检验 20 个为真的统计假设，那么至少错误地拒绝了一个为真的零假设的最大概率为 $20 \times 0.01 = 0.20$，即五个为真的假设检验中就有一个被当做错误假设被拒绝。这提醒我们，有时天真的“发掘数据”可能会

导致严重的错误。

随机变量

随机变量是定义在样本空间上取值为实数的函数。对于之前所提及的样本空间 $S = \{HH, HT, TH, TT\}$，一个记录掷硬币结果为正面的随机变量 X 可定义为：

结果	X 的取值
HH	2
HT	1
TH	1
TT	0

对于此例，如果 X 为离散随机变量，那么，我们通常把 $\Pr(X = x)$ 写成 $p(x)$，其中，大写字母 X 代表随机变量，小写字母 x 表示变量的特殊值。[26]例如，掷硬币实验的四个结果发生的概率均为 0.25，那么，出现正面的概率分布为：

	x	$p(x)$
$\{TT\}\rightarrow$	0	0.25
$\{HT, TH\}\rightarrow$	1	0.50
$\{HH\}\rightarrow$	2	0.25
	总计	1.00

该表记录了所有事件匹配到每个随机变量 x 值后的结果。

一个随机变量 X 的累积分布函数 CDF 给出可观测到的变量值小于或者等于某个特殊值的概率，记做 $P(x)$：

$$P(x) = \Pr(X \leqslant x) = \sum_{x' \leqslant x} p(x')$$

对于上述例子：

x	$p(x)$
0	0.25
1	0.75
2	1.00

如果随机变量是在一个连续变量空间中定义的，那么，这些随机变量本身也可能是连续的。这里，我们仍然用$p(x)$代表 $\Pr(X \leqslant x)$，但是，对于随机变量 X 的每一个具体值来说[27]，这种表示就会显得毫无意义。概率密度函数 $p(x)$是离散概率分布的连续模拟，定义为 $p(x) \equiv dP(x)/dx$。[28]变换后得到[29]：

$$P(x) = \int_{-\infty}^{x} p(x)dx$$

$$\Pr(x_0 \leqslant X \leqslant x_1) = P(x_1) - P(x_0) = \int_{x_0}^{x_1} p(x)dx$$

因此，如图 3.1 所示，密度函数以下的区域代表概率。[30]

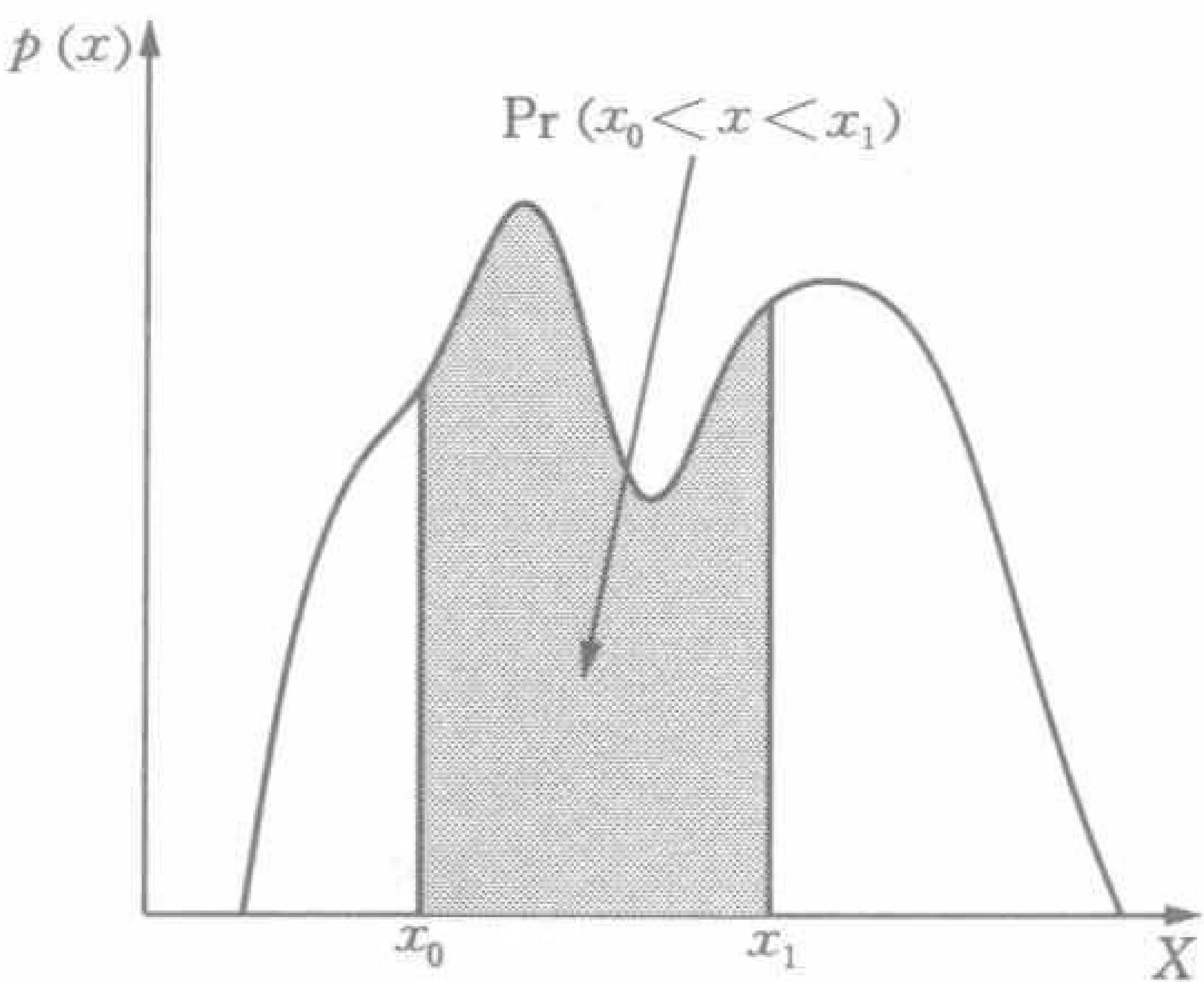


图 3.1 概率密度曲线 $p(x)$以下的区域为概率

最简单的连续概率分布是均匀分布：

$$p(x) = \begin{cases} 0 & a > x \\ \dfrac{1}{b-a} & a \leqslant x \leqslant b \\ 0 & x > b \end{cases} \qquad [3.6]$$

其密度函数见图 3.2(a)，相应的累积分布函数见图 3.2(b)。密度函数下方整个区域的大小为 1。这里，

$$\int_{-\infty}^{\infty} p(x)dx = \int_{a}^{b} p(x)dx = \frac{1}{b-a}(b-a) = 1$$

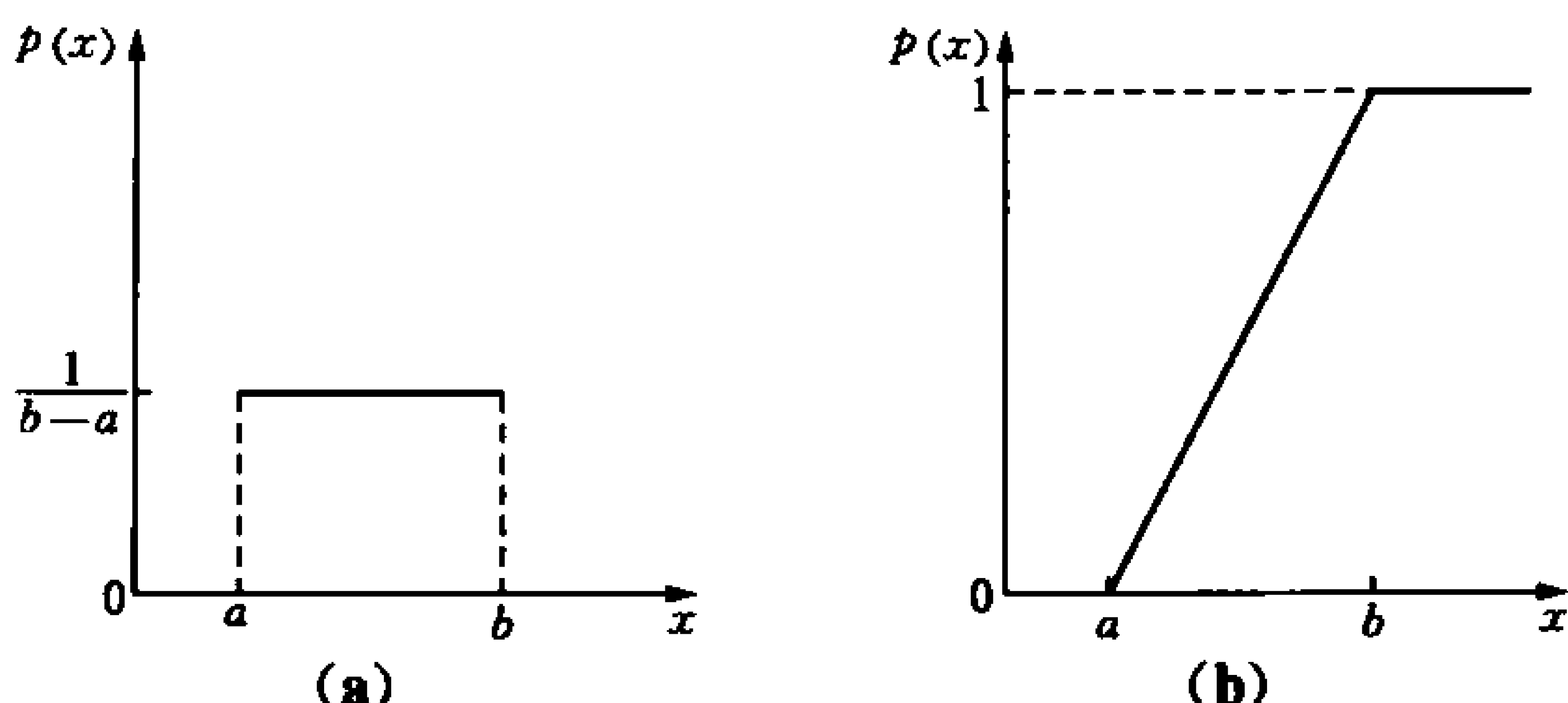


图 3.2 (a)均匀分布的概率密度函数 $p(x)$和(b)均匀分布的累积分布函数 $p(x)$

一个随机变量的支持是一组概率或者概率密度不为 0 的数值。因此，均匀分布的支持为 $a \leqslant X \leqslant b$。

随机变量的两个基本属性是其期望值(或平均值)和方差。[31]从期望值可以知道随机变量概率分布的中心(这道理就如同一系列取值的均值指明了由这些取值所构成的分布的中心)，方差记录了分布相对于期望值的分散程度。随机变量的期望值为随机变量通过多次重复试验得到的取值的均值，方差为取值和期望值之间的均方距离。

对于离散事件，随机变量 X 的期望值记做 $E(X)$或者 μ_X，表示为：

$$E(X) \equiv \sum_{\text{all}\,x} xp(x)$$

对于连续事件，随机变量 X 的期望值表示为[32]：

$$E(X) \equiv \int_{-\infty}^{\infty} xp(x)dx$$

一个随机变量 X 的方差记做 $V(X)$ 或者 σ_X^2，定义为：

$$V(X) \equiv E[(X-\mu_X)^2]$$
$$= E(X^2) - \mu_X^2$$

因此，对于离散事件，

$$V(X) \equiv \sum_{\text{all}X'} (x-\mu_X)^2 p(x)$$

那么，对于连续事件，

$$V(X) \equiv \int_{-\infty}^{\infty} (x-\mu_X)^2 p(x) dx$$

随机变量的方差是用平方单位来表示的（例如，“出现正面的次数的平方”），但是标准差 $\sigma \equiv + \sqrt{\sigma^2}$ 的量度单位与变量相同。

对于我们的例子，

x	$p(x)$	$xp(x)$	$x-\mu$	$(x-\mu_X)^2 p(x)$
0	0.25	0.00	−1	0.25
1	0.50	0.50	0	0.00
2	0.25	0.50	1	0.25
总计	1.00	μ=1.00		σ=0.50

因此，$E(X)=1$，$V(X)=0.5$，$\sigma=\sqrt{0.5}\approx 0.707$。同样，对于均匀分布（方程 3.6），

$$E(X) = \int_b^b x\left(\frac{1}{b-a}\right)dx = \frac{a+b}{2}$$

$$V(X) = \int_b^b \left(x-\frac{a+b}{2}\right)^2 \left(\frac{1}{b-a}\right)dx = \frac{(a-b)^2}{12}$$

两个离散随机变量 X_1 和 X_2 的联合概率分布提供了同时

观测到两个变量的任意一对取值的概率。我们把 $\Pr(X_1 = x_1$ 和 $X_2 = x_2)$ 记做 $p_{12}(x_1, x_2)$。但是 p 的下标时常会引起歧义，因此，我们将其简化为 $p(x_1, x_2)$。两个连续变量的 $p(x_1, x_2)$ 的联合概率分布与离散变量的定义类似。多个随机变量的联合概率分布的表示方法为 $p(x_1, x_2, \cdots, x_n)$。

不同于随机变量的联合概率分布，$p_1(x_1)$ 为随机变量 X_1 的边缘概率分布或者边缘概率密度。其中，$p_1(x_1) = \sum_{x_2} p(x_1, x_2)$ 或者 $p_1(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_2$，我们常常忽略下标而将其记做 $p(x_1)$。

在一个掷硬币实验中，我们用 X_1 记录出现正面的次数，并定义 $X_2 = 1$ 时，两次掷硬币得到的结果相同，$X_2 = 0$ 时，两次掷硬币的结果不同，那么，

结果	Pr	x_1	x_2
HH	0.25	2	1
HT	0.25	1	0
TH	0.25	1	0
TT	0.25	0	1

随机变量 X_1 和 X_2 的联合边缘分布如下表所示：

$p(x_1, x_2)$			
x_1	x_2		$p(x_1)$
	0	1	
0	0	0.25	0.25
1	0.50	0	0.50
2	0	0.25	0.25
$p(x_2)$	0.50	0.50	1.00

给定 X_2，X_1 的条件概率或者条件概率密度为：

$$p_{1|2}(x_1 \mid x_2) = \frac{p_{12}(x_1, x_2)}{p_2(x_2)}$$

与之前相同,为方便起见,我们常常会省略下标,记做 $p(x_1|x_2)$。对于该实验,当 $x_2=1$ 时和当 $x_2=0$ 时,条件概率 $p(x_1|x_2)$为：

$p(x_1\|x_2)$		
x_1	x_2	
	0	1
0	0	0.5
1	1.0	0
2	0	0.5
总计	1.0	1.0

给定 $X_2=x_2$ 时,将 X_1 的条件期望值记做 $E_{1|2}(X_1|x_2)$或者 $E(X_1|x_2)$,它是从条件分布 $p_{1|2}(x_1|x_2)$而来的。同样,给定 $X_2=x_2$ 时,X_1 的条件方差记做 $V_{1|2}(X_1|x_2)$或者 $V(X_1|x_2)$。对于一个离散事件,

$$E_{1|2}(X_1 \mid x_2) = \sum_{x_1} x_1 p_{1|2}(x_1 \mid x_2)$$

$$V_{1|2}(X_1 \mid x_2) = \sum_{x_1} [x_1 - E_{1|2}(X_1 \mid x_2)]^2 p_{1|2}(x_1 \mid x_2)$$

将具体数值代入后,得到：

$$E_{1|2}(X_1 \mid 0) = 0(0)+1(1)+0(2) = 1$$

$$V_{1|2}(X_1 \mid 0) = 0(0-1)^2+1(1-1)^2+0(2-1)^2 = 0$$

$$E_{1|2}(X_1 \mid 1) = 0.5(0)+0(1)+0.5(2) = 1$$

$$V_{1|2}(X_1 \mid 0) = 0.5(0-1)^2+0(1-1)^2+0.5(2-1)^2 = 1$$

如果对于随机变量 X_1 和 X_2 的任意取值,都有 $p(x_1)=$

$p(x_1 \mid x_2)$，那么，我们说 X_1 和 X_2 是独立随机变量。也就是说，如果 X_1 和 X_2 为独立随机变量，那么，X_1 的条件分布与边缘分布是等价的。对于以上题设，其独立性的等价条件还有 $p(x_2) = p(x_2 \mid x_1)$，$p(x_1, x_2) = p(x_1)p(x_2)$，当 X_1 和 X_2 为独立随机变量时，它们的联合概率或者概率密度是它们边缘概率或概率密度的乘积。在此例中，X_1 和 X_2 明显不是独立随机变量。推广之，对于包含 n 个随机变量的独立集合 $\{X_1, X_2, \cdots, X_n\}$，其每个子集 $\{X_a, X_b, \cdots, X_m\}(m \geqslant 2)$ 有：

$$p(x_a, x_b, \cdots, x_m) = p(x_a)p(x_b) \cdots p(x_m)$$

两个随机变量的协方差为它们是否线性独立的量度：

$$C(X_1, X_2) = \sigma_{12} \equiv E[(X_1 - \mu_1)(X_2 - \mu_2)]$$
$$= E(X_1X_2) - \mu_1\mu_2$$

当随机变量 X_1 较大的取值与随机变量 X_2 较大的取值相关时，其协方差为正；当随机变量 X_1 较大的取值与随机变量 X_2 较小的取值相关时，其协方差为负（反之亦然）；当两个随机变量属于独立随机变量时，协方差为 0，但是随机变量的独立性并不是协方差为 0 的充分必要条件，即两个随机变量可呈现非线性相关，此时协方差仍可为 0。在之前的例子中，X_1 和 X_2 并不是独立随机变量，但是 σ_{12} 无疑为 0（读者自己可以证实）。变量本身的协方差就是其本身的方差：$C(X, X) = V(X)$。

随机变量 X_1 和 X_2 的相关性 $\rho_{12} \equiv \sigma_{12}/\sigma_1\sigma_2$ 是个标准化后的协方差。相关性的最小值 $\rho = -1$，它表示随机变量之间存在完美的反线性关系。同样，相关性的最大取值是 $\rho = 1$，它表示随机变量之间存在完美的正线性关系。当 $\rho = 0$ 时，

协方差为 0，此时随机变量间不存在线性关系。

为方便起见，我们常常将一系列随机变量写成一个随机向量。例如 $\underset{(n\times1)}{\mathbf{x}}=[X_1, X_2, \cdots, X_n]'$。一个随机向量的期望值就是其中元素的期望值组成的向量，记做：

$$E(x)=\mu_x\equiv[E(X_1), E(X_2), \cdots, E(X_n)]'$$

随机向量

x 的方差—协方差矩阵定义与纯量方差类似，表达式为：

$$V(\mathbf{x})=\underset{(n\times n)}{\Sigma_{\mathbf{xx}}}\equiv E[(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})']=\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2\end{bmatrix}$$

$V(\mathbf{x})$的对角元是变量 X 的方差，非对角元是其协方差。方差—协方差矩阵$V(\mathbf{x})$是一个对称半正定矩阵。两个随机向量 $\underset{(n\times1)}{\mathbf{x}}$ 和 $\underset{(n\times1)}{\mathbf{y}}$ 的协方差矩阵表示为：

$$C(\mathbf{x}, \mathbf{y})=\underset{(n\times m)}{\boldsymbol{\Sigma}_{\mathbf{xy}}}\equiv E[(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})(\mathbf{y}-\boldsymbol{\mu}_\mathbf{y})']$$

$$=\begin{bmatrix}\sigma_{x_1y_1} & \sigma_{x_1y_2} & \cdots & \sigma_{x_1y_m}\\ \sigma_{x_2y_1} & \sigma_{x_2y_2} & \cdots & \sigma_{x_2y_m}\\ \vdots & \vdots & & \vdots\\ \sigma_{x_ny_1} & \sigma_{x_ny_2} & \cdots & \sigma_{x_ny_m}\end{bmatrix}$$

其包括了所有随机向量 X 和 Y 内所有元素的所有对协方差。

随机变量的变换

假设随机变量 Y 是随机变量 X 的线性函数 $a+bX$（其

中，a、b 为常数），X 的期望值和方差分别为 μ_X 和 σ_X^2，那么，

$$E(Y)=\mu_Y=\sum_x(a+bx)p(x)$$

$$=a\sum p(x)+b\sum xp(x)$$

$$=a+b\mu_X$$

$$V(Y)=E[(Y-\mu_Y)^2]=E\{[(a+bX)-(a+b\mu_X)]^2\}$$

$$=b^2E[(X-\mu_X)^2]=b^2\sigma_X^2$$

现在，假设 Y 是两个随机变量 X_1 和 X_2 的线性函数 $a_1X_1+a_2X_2$，X_1 和 X_2 所对应的期望值分别为 μ_1、μ_2，方差分别为 σ_1^2、σ_2^2，协方差为 σ_{12}。那么，我们得到：

$$E(Y)=\mu_Y=\sum_{x_1}\sum_{x_2}(a_1x_1+a_2x_2)p(x_1, x_2)$$

$$=\sum_{x_1}\sum_{x_2}a_1x_1p(x_1, x_2)+\sum_{x_1}\sum_{x_2}a_2x_2p(x_1, x_2)$$

$$=a_1\sum_{x_1}x_1p(x_1)+a_2\sum_{x_2}x_2p(x_2)$$

$$=a_1\mu_1+a_2\mu_2$$

$$V(Y)=E[(Y-\mu_Y)^2]$$

$$=E\{[(a_1x_1+a_2x_2)-(a_1\mu_1+a_2\mu_2)]^2\}$$

$$=a_1^2E[(X_1-\mu_1)^2]+a_2^2E[(X_2-\mu_2)^2]$$

$$+2a_1a_2E[(X_1-\mu_1)(X_2-\mu_2)]$$

$$=a_1^2\sigma_1^2+a_2^2\sigma_2^2+2a_1a_2\sigma_{12}$$

其中，X_1 和 X_2 是独立随机变量，因此 $\sigma_{12}=0$，那么，以上表

达式可简化为 $V(Y) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2$。

连续事件的规则与离散事件相同。例如,如果 $Y = a + bx$ 是一个连续变量 X 的线性函数,那么[33],

$$E(Y) = \int_{-\infty}^{\infty}(a + bx)p(x)dx$$

$$= a\int_{-\infty}^{\infty}p(x)dx + b\int_{-\infty}^{\infty}xp(x)dx$$

$$= a + bE(X)$$

随机向量的变换

将以上结论推广到随机向量中后,我们得到:如果 $\underset{(m\times 1)}{\mathbf{y}}$ 是随机向量 $\mathbf{x}$ 的线性变换 $\underset{(m\times n)}{\mathbf{A}}\ \underset{(n\times 1)}{\mathbf{x}}$,随机向量 $\mathbf{x}$ 的期望值是 $E(x) = \mu_X$,方差—协方差矩阵为 $V(\mathbf{x}) = \Sigma_{xx}$,则有:

$$E(\mathbf{y}) = \boldsymbol{\mu}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\mu}_{\mathbf{x}}$$

$$V(\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{yy}} = \mathbf{A}\boldsymbol{\Sigma}_{xx}A'$$

如果随机向量 $\mathbf{x}$ 的元两两独立,那么,所有的非对角元都为 0,$\mathbf{y}$ 中每个元的方差可简单表示为:

$$\sigma_{Y_i}^2 = \sum_{j=1}^{n} a_{ij}^2\sigma_{X_j}^2$$

有时,对于 $\mathbf{y} = f(\mathbf{x})$,我们需要知道的不仅是 $E(\mathbf{y})$ 和 $V(\mathbf{y})$,还有 $\mathbf{y}$ 的概率分布。而且,变换操作 $f(\cdot)$ 也有可能为非线性操作。假设 $\mathbf{y}$ 和 $\mathbf{x}$ 中的元素数目相等(均为 n)、f 函数是可微的、f 与 $\mathbf{x}$ 的范围内的值是一一对应的(每一个 $\mathbf{x}$ 都对应一个唯一的 $\mathbf{y}$),且最后一个属性暗示了该函数有反函数 $\mathbf{x} = f^{-1}(\mathbf{y})$。那么,$\mathbf{y}$ 的概率密度可表示为:

$$p(\mathbf{y}) = p(\mathbf{x}) \left| \det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right) \right| = p(\mathbf{x}) \left| \det\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right) \right|^{-1}$$

其中，$|\det(\partial \mathbf{x}/\partial \mathbf{y})|$叫做“雅可比迭代”，它是 $(n \times n)$ 行列式的绝对值：

$$\det \begin{bmatrix} \frac{\partial X_1}{\partial Y_1} & \cdots & \frac{\partial X_n}{\partial Y_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial X_1}{\partial Y_n} & \cdots & \frac{\partial X_n}{\partial Y_n} \end{bmatrix}$$

$|\det(\partial \mathbf{y}/\partial \mathbf{x})|$的定义与$|\det(\partial \mathbf{x}/\partial \mathbf{y})|$类似。

第2节 | 离散概率分布

在本章节，我主要对一些重要的离散概率分布类进行详解，如二项分布与伯努利分布、多项分布、泊松分布（该分布可构建出近似二项分布），还有负二项分布。我们所说的概率分布（例如，二项分布）其实是一个类，但为方便起见，我们只说二项分布。本章节的有关离散分布的内容和之后连续分布的内容均在统计推理和统计建模中扮演着非常重要的角色。

二项分布和伯努利分布

前文提到的掷硬币实验引出了一个二项分布随机变量，该变量记录了一个硬币两次投掷后得到正面的次数。将此例引申后，我们让随机变量 X 记录一个硬币 n 次投掷后得到正面的次数。其中，π 表示任意投掷得到正面的概率（不一定为 0.5），$1-\pi$ 则为得到反面的概率。[34]那么，观测到 x 个正面和 $n-x$ 个反面的情况可用一个二项分布来表示：

$$p(x) = \binom{n}{x}\pi^{x}(1-\pi)^{n-x} \qquad [3.7]$$

其中，x 是 0 到 n 的任意整数，因子 $\pi^{x}(1-\pi)^{n-x}$ 是在特定情

况下观测到 x 个正面和 $n-x$ 个反面的概率。$\binom{n}{x} \equiv n!/[x!(n-x)!]$ 是二项系数，它是出现 x 个正面和 $n-x$ 个反面的所有组合的数量。[35]

二项分布随机变量 X 的期望值为 $E(X)=n\pi$，方差为 $V(X)=n\pi(1-\pi)$。图 3.3 展示了当 $n=10$ 及 $\pi=0.7$ 时的二项分布。如果乘积项 $n\pi$ 和 $n(1-\pi)$ 足够大(例如，都至少等于 10)，那么离散二项分布可以近似看做连续正态分布，且其均值和标准差都与连续正态分布相同。

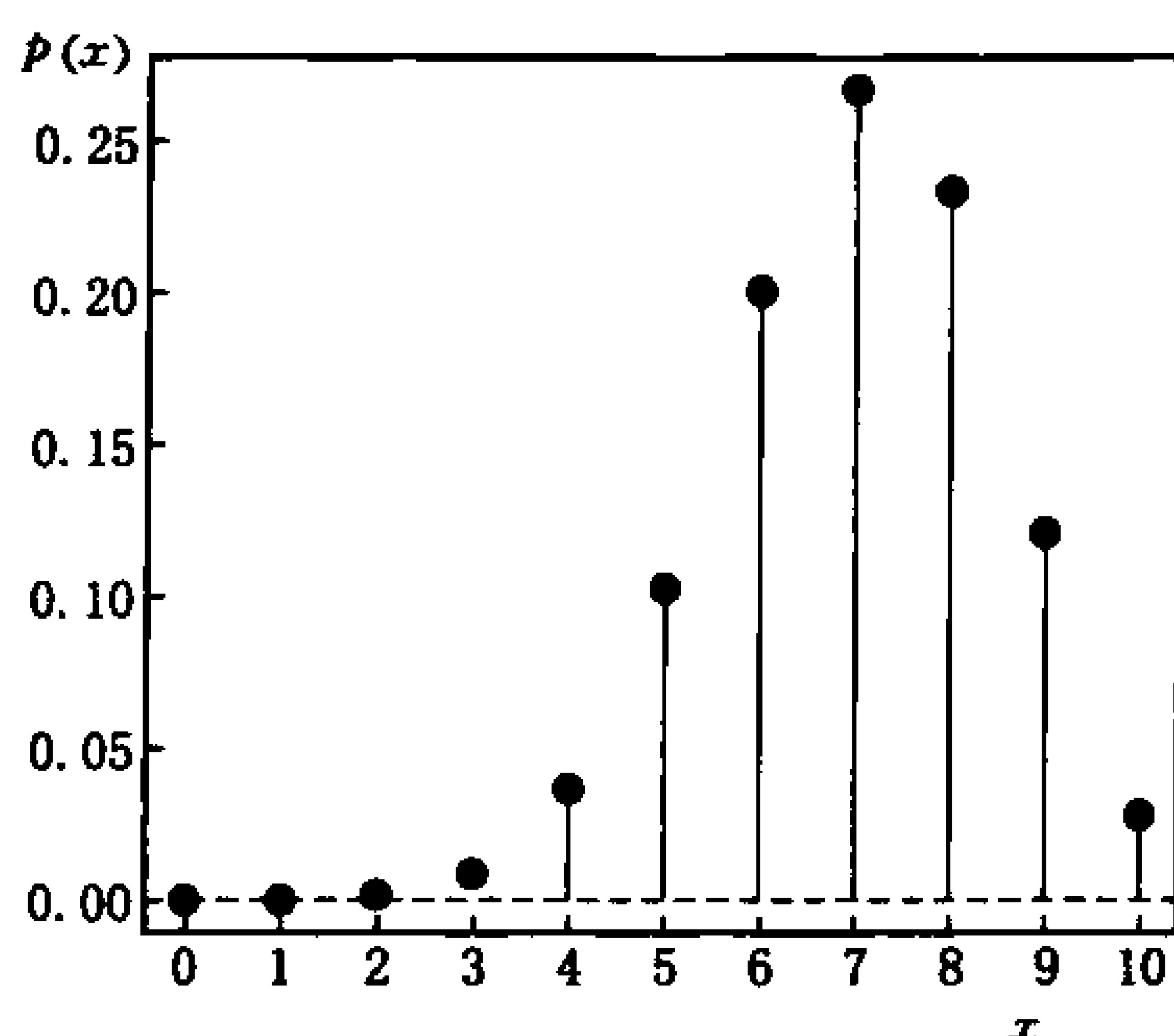


图 3.3 $n=10$ 及 $\pi=0.7$ 时的二项分布

二项分布随机变量与伯努利随机变量不同，后者在取 0 和 1 的时候所对应的概率分别是 π 和 $1-\pi$，其均值和方差分别为 $E(X)=\pi$，$V(X)=\pi(1-\pi)$。伯努利随机变量可以用来对一次投掷建模。例如，假设 $X=1$ 为出现硬币的正面，$X=0$ 为出现硬币的反面，那么，独立且同分布的伯努利变量的加和是一个二项分布。

多项分布

假设在 n 次重复独立的实验中，每一次实验的结果都出

现在k个不同的结果类别中(对于该实验,总共会出现k种结果)。我们让随机变量X_i表示类别i中的结果数量,让π_i表示每次实验的结果落入类别i中的概率。那么,$\sum_{i=1}^{k}\pi_i=1$,$\sum_{i=1}^{k}X_i=n$。

如果我们掷n次骰子,让X_1记录出现1的次数,X_2记录出现2的次数……X_6记录出现6的次数。因此,$k=6$,π_1表示掷出1的概率,π_2表示掷出2的概率,等等。如果骰子就是普通的骰子,即其各个面的数字不同,则有$\pi_1=\pi_2=\cdots=\pi_6=1/6$。

推广到一般情况可知,如果向量随机变量$\mathbf{x}\equiv[X_1, X_2, \cdots, X_k]'$符合多项分布,则有:

$$p(\mathbf{x})=p(x_1, x_2, \cdots, x_k)=\frac{n!}{x_1!x_2!\cdots x_k!}\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}$$

该公式的原理与二项分布公式相似,即$\pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}$分别为在特定情况下,结果在类别1中出现的概率,结果在类别2中出现的概率,等等。$n!/(x_1!\ x_2!\ \cdots x_k!)$记录了不同组合的个数。如果$k=2$,那么,$x_2=n-x_1$,此时,多项分布即简化为二项分布(见方程3.7)。

随机向量$\mathbf{x}$中,元素的期望值为$E(X_i)=n\pi_i$,方差为$V(X_i)=n\pi_i(1-\pi_i)$,其对应的协方差为$C(X_i, X_j)=-n\pi_i\pi_j$。

泊松分布

19世纪法国数学家西蒙-丹尼·泊松(Siméon-Denis Poisson)引入了以其名命名的一个近似二项分布。该近似在n足够大、π足够小且其乘积$\lambda\equiv n\pi$适中的情况下成立。泊

松分布的表达式为：

$$p(x)=\frac{\lambda^{x}e^{-\lambda}}{x!}(x=0,1,2,3,\cdots \text{且} \lambda>0)$$

尽管 X 所包含的均是非负整数，但由于 $p(x)\approx 0$，因此，该近似只有在 x 足够大的情况下才可行（这里，e 是一个数学常数，$e\approx 2.71828$）。

泊松分布只用于极少见或不经常发生的现象。假设我们所观测到的过程所产生的事件比较特殊（如出生或者其他自发事件），对于事件 X，我们会记录下其在某个固定时间段发生的次数，如果该发生次数符合以下条件，则其遵循泊松分布：(1)尽管事件发生的时间是随机的，但是在某个观测间隔下，其发生率是固定的。(2)如果我们将注意力放在一个充分小、间隔长度为 s 的子间隔内，那么，在该间隔内观测到一个事件的概率与其所在的间隔长度 λs 成正比，在该间隔内观测到多于一个时间的概率几乎小到可以忽略。这样，参数 λ 即事件的发生率。(3)在不重叠子区间发生的事件是独立事件。

泊松随机变量的期望值是 $E(X)=\lambda$，其方差 $V(X)$ 也是 λ。图 3.4 描述了参数 $\lambda=5$（有五个事件发生在观测的固定区间）时的泊松分布。

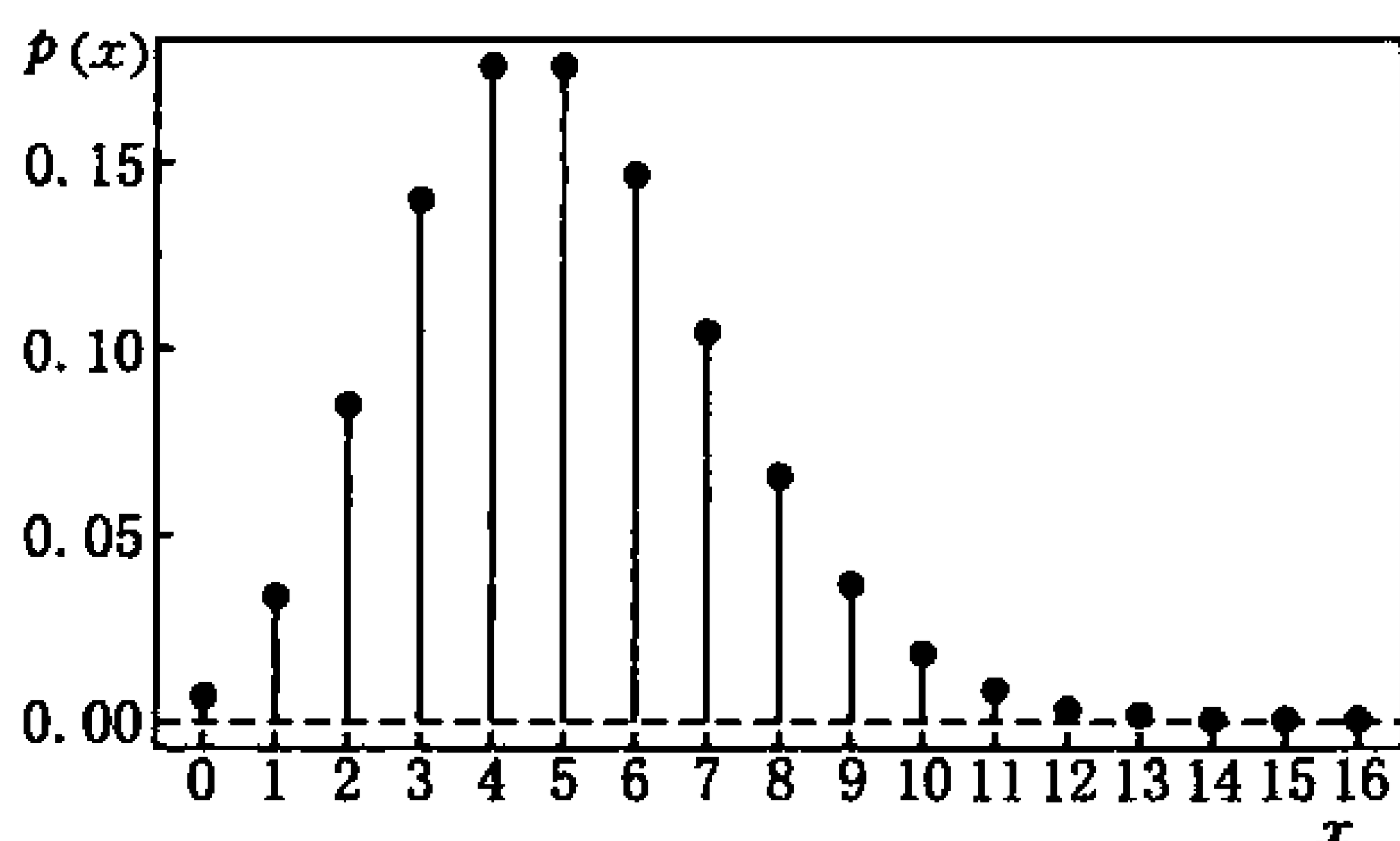


图 3.4　参数为 $\lambda=5$ 的泊松分布

负二项分布

假设在掷硬币实验中，每次投掷都是独立的，并一直持续到一个目标数量，如出现 s 个正面后停止，此时，我们让随机变量 X 记录目标数量达到前，出现反面的次数。那么，X 遵循一个负二项分布，其概率分布的表达式为：

$$p(x)=\binom{s+x-1}{x}\pi^{s}(1-\pi)^{x}\quad(x=0,\ 1,\ 2,\ \cdots)$$

其中，π 是每次掷硬币出现正面的概率，该负二项分布的期望值为 $E(X)=s(1-\pi)/\pi$，方差为 $V(X)=s(1-\pi)/\pi^{2}$。图 3.5 表示当 $s=4$ 及 $\pi=0.5$ 时的负二项分布。

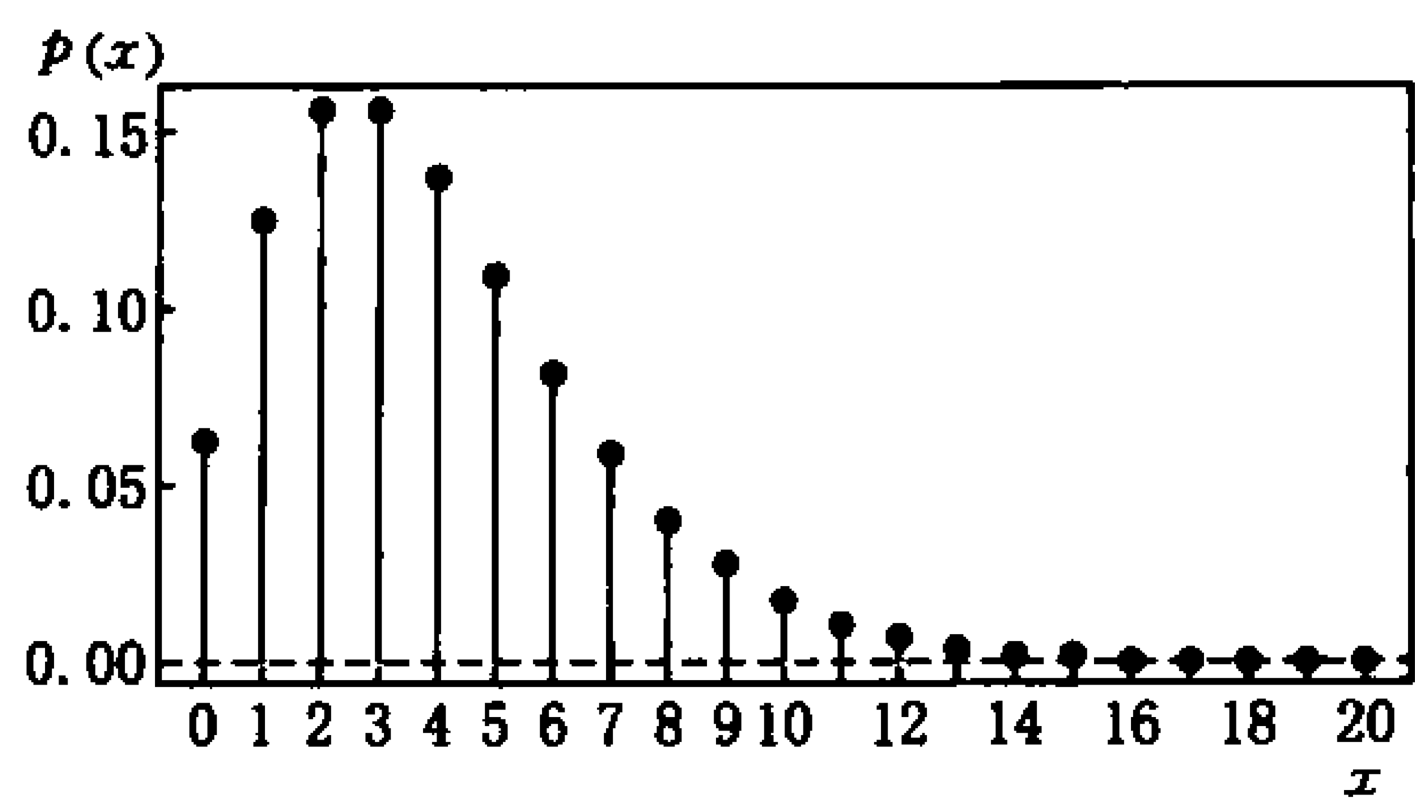


图 3.5　$s=4$ 及 $\pi=0.5$ 时的负二项分布

第 3 节 | 连续分布

在本章节中，我会介绍一些重要的连续分布类型，如正态分布、卡方分布、t 分布、F 分布、多元正态分布、指数分布、逆高斯分布、γ 及 β 分布。

正态分布

正态分布(或高斯分布)随机变量 X 的概率密度函数为：

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (-\infty < x < \infty)$$

其中，分布参数 μ 和 σ^2 分别为 X 的均值和方差。因此，对于每个 μ 和 σ^2，都有一个不同的正态分布。图 3.6 给我们列出了几个例子。正态分布常见的缩写形式为 $X \sim N(\mu, \sigma^2)$，它表示 X 是一个以 μ 为均值、以 σ^2 为方差的正态分布。[36] 尽管法国数学家亚伯拉罕·棣莫弗(Abraham de Moivre)已于1973 年第一次引入了这个近似二项分布的概念，但是高斯分布仍是以伟大的德国数学家卡尔·弗里德里希·高斯(Carl Friedrich Gauss)这一对正态分布有着重要贡献的数学家命名的。

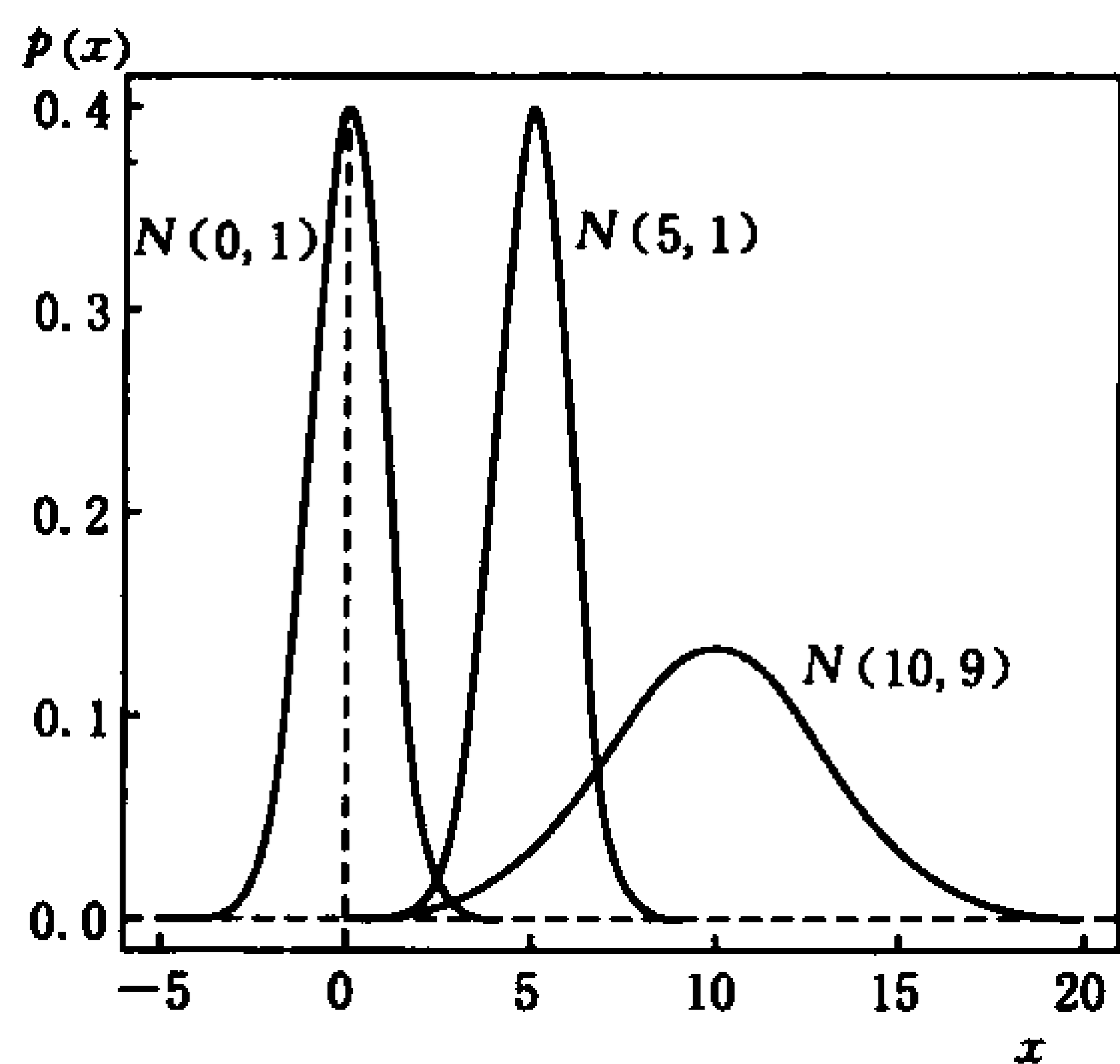


图 3.6　正态概率函数：$N(0, 1)$、$N(5, 1)$ 和 $N(10, 9)$

单位正态分布(或者标准正态分布)的随机变量 $Z \sim N(0, 1)$ 的密度函数在统计上有着非常重要的用途，其表达式为：

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \quad (-\infty < z < \infty)$$

该分布的累积分布函数 $\Phi(z)$ 如图 3.7 所示。任意正态分布随机变量 $X \sim N(\mu, \sigma^2)$ 都可以转化为标准形式[37]：

$$Z \equiv \frac{X-\mu}{\sigma}$$

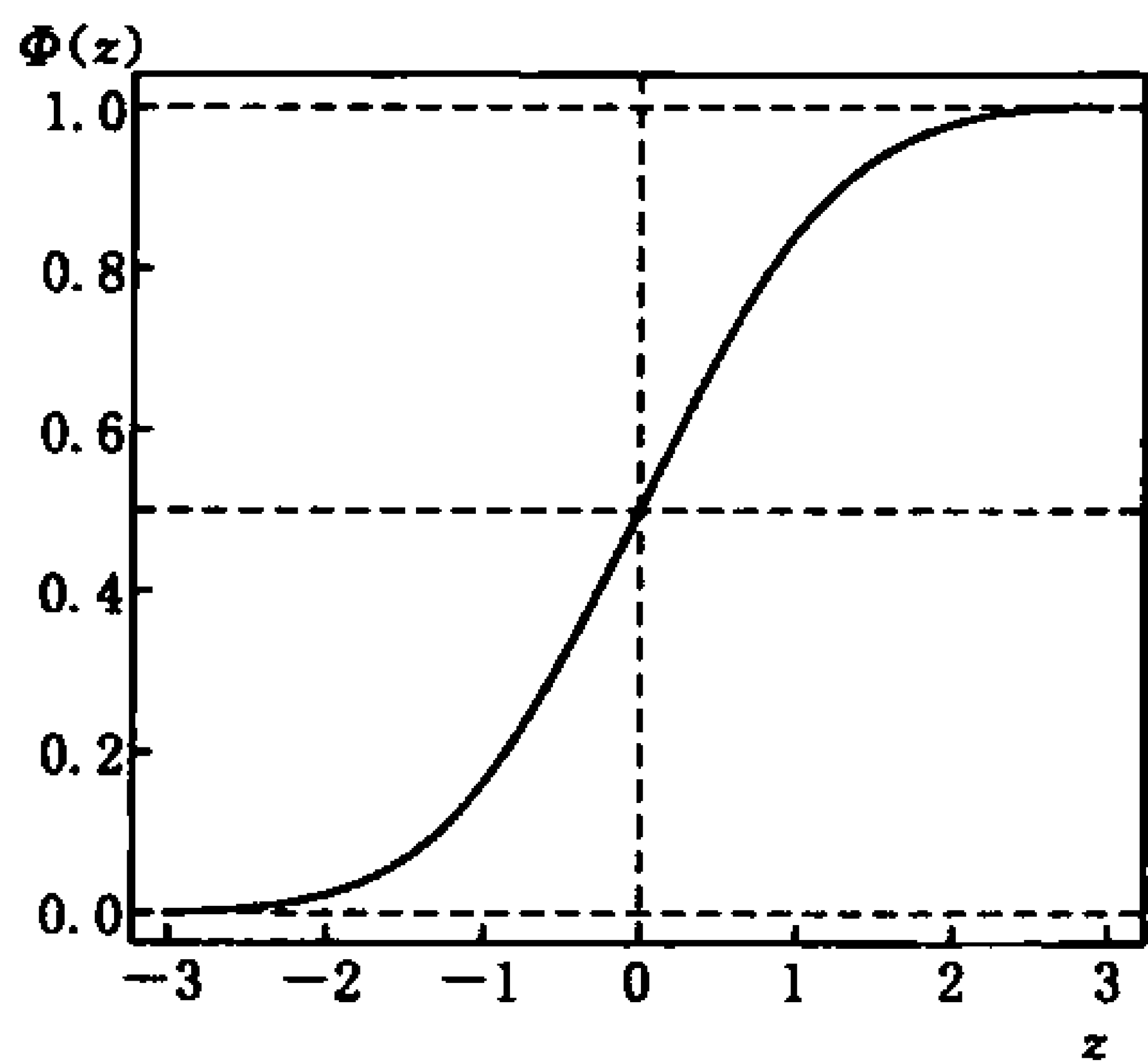


图 3.7　单位正态分布的累积分布函数 $\Phi(z)$

卡方分布

如果 $Z_1, Z_2, \cdots, Z_n$ 为独立的标准正态分布随机变量，那么，

$$X^2 \equiv Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

其遵循一个含有 n 个自由度的卡方分布，简写为 χ_n^2。卡方随机变量的概率密度函数为：

$$p(x^2) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}(x^2)^{(n-2)/2}\exp(-x^2/2) \quad (x^2 \geqslant 0)$$

其中，$\Gamma(\cdot)$是一个 γ 分布函数，

$$\Gamma(\upsilon) \equiv \int_0^{\infty} e^{-\upsilon} z^{\upsilon-1} dz \quad (\text{通用参数 } \upsilon > 0) \qquad [3.8]$$

它是连续阶乘函数的一般形式。特别是当 υ 等于一个非负整数时，$\upsilon! = \Gamma(\upsilon+1)$，我们有：

$$\Gamma\left(\frac{n}{2}\right) = \begin{cases} \left(\frac{n}{2}-1\right)! & (n\text{ 为偶数}) \\ \left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\sqrt{\pi} & (n\text{ 为奇数}) \end{cases}$$

卡方随机变量的期望值和方差分别为 $E(X^2) = n$ 和 $V(X^2) = 2n$。图 3.8 列出了一些卡方分布。如图所示，卡方分布是正偏的，但是随着自由度的增加，该分布变得越来越对称，即趋近正态分布。

如果 $X_1^2, X_2^2, \cdots, X_k^2$ 分别为自由度是$n_1, n_2, \cdots, n_k$ 的卡方随机变量，那么 $X \equiv X_1^2 + X_2^2 + \cdots + X_n^2$ 遵循自由度为

$n = n_1 + n_2 + \cdots + n_k$ 的卡方分布。

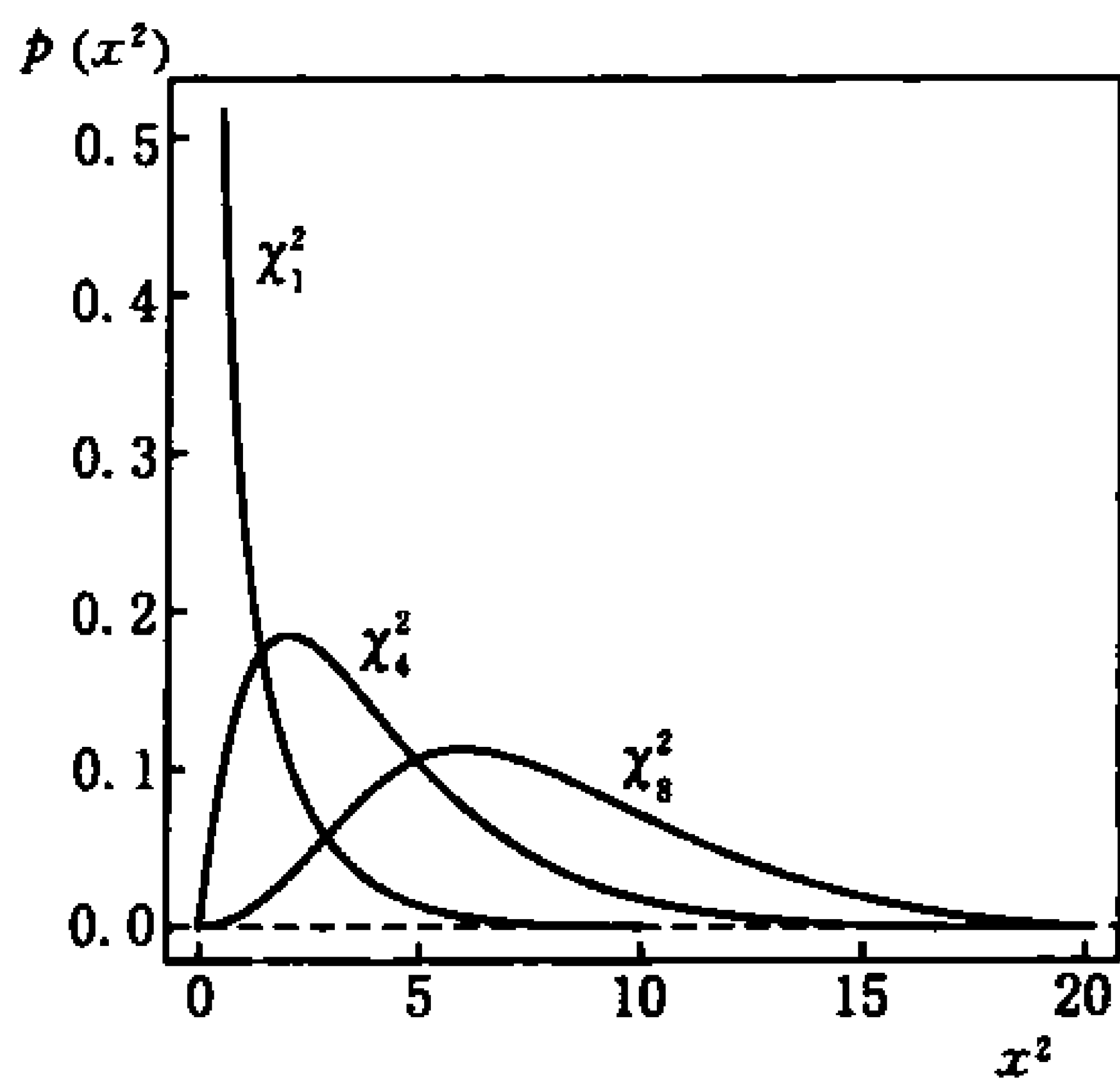


图 3.8 卡方密度函数：χ_1^2、χ_4^2 和 χ_8^2

学生 t 分布

如果 Z 遵循标准正态分布，且 X^2 遵循 n 个自由度的卡方分布，那么，

$$t \equiv \frac{Z}{\sqrt{\frac{X^2}{n}}}$$

这就是一个有 n 个自由度的学生 t 随机变量，简写为 t_n。[38] 其概率密度函数为：

$$p(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \times \frac{1}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}} \quad (-\infty < t < \infty)$$

[3.9]

该公式在 $t = 0$ 点中心对称，因此 $E(t) = 0$。[39]我们可以发现，对于任意 $n > 2$，$V(t) = n/(n-2)$，因此，对于自由度较

小的分布，t 的方差比较大，随着 n 的增加，方差越来越趋近 1。

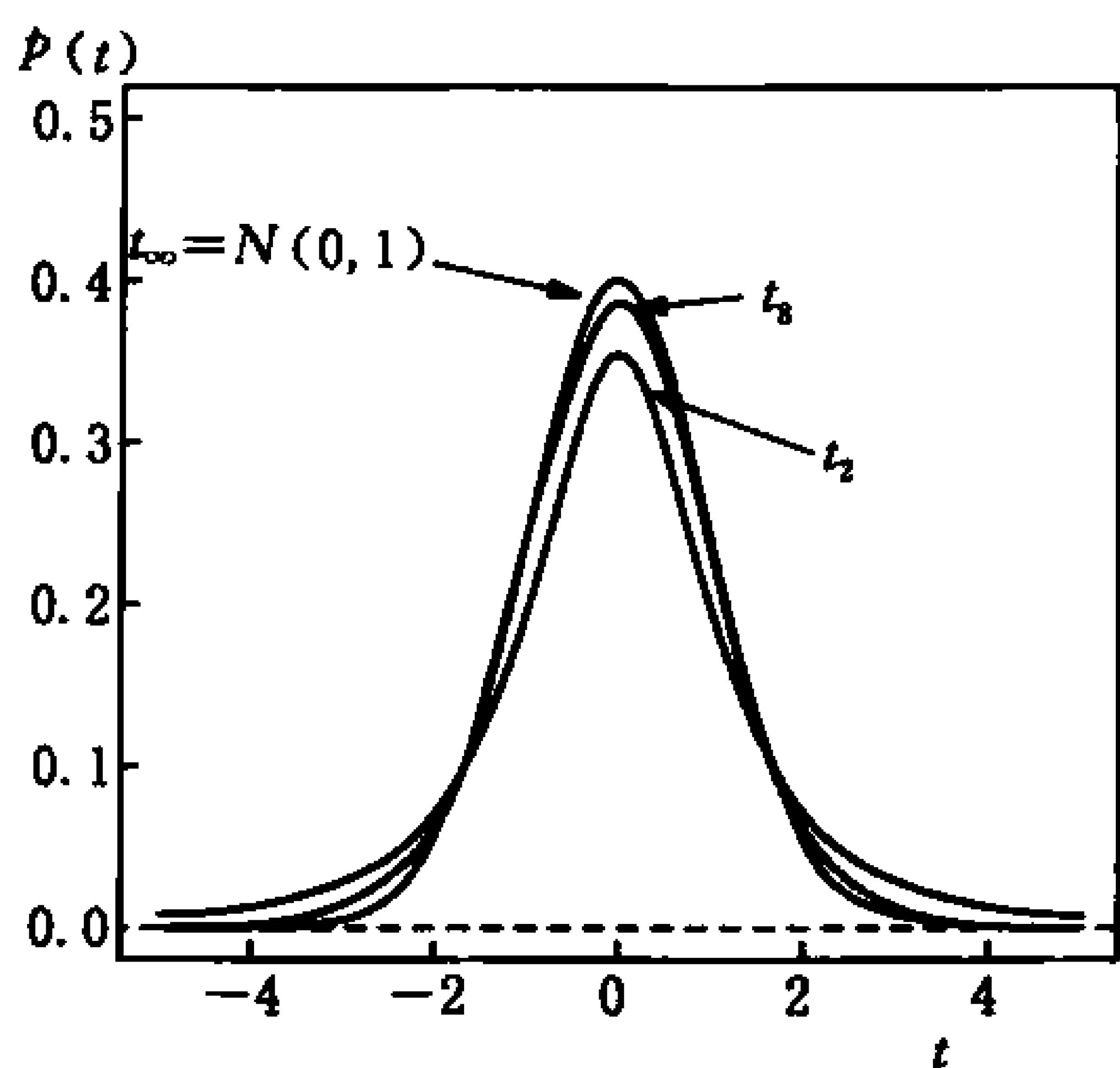


图 3.9 t 密度函数：t_2、t_8 及 $t_\infty = N(0, 1)$

图 3.9 描述了一些 t 分布图形。当自由度增加时，t 分布越来越趋近标准正态分布，直到 $t_\infty = N(0, 1)$。当 $n \geqslant 30$ 时，t 分布的方差就趋近于 1，t 分布也就可以近似看做标准正态分布。

学生 t 分布以 20 世纪英国都柏林吉尼斯啤酒厂的一名统计学家威廉・西利・戈塞特(William Sealy Gossett)命名。戈塞特曾以“学生”为笔名在《生物计量学》杂志上发表了论文《平均数的规律误差》。这篇论文开创了小样本统计理论的先河。学生 t 分布对小样本统计推理的发展起到了举足轻重的作用。

F 分布

令 X_1^2 和 X_2^2 分别代表自由度为 n_1 和 n_2 的卡方随机变量。那么，

$$F \equiv \frac{X_1^2/n_1}{X_2^2/n_2}$$

它遵循自由度为 n_1 和 n_2 的 F 分布，简写为 F_{n_1,n_2}。F 分布是美国统计学家乔治·W. 斯内德克(Geroge W. Snedecor)为奖励其发现者——伟大的英国统计学家 R. A. 费希尔爵士(Sir R. A. Fisher)而命名的。

F 分布的概率密度为：

$$p(f)=\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{n_1/2}\cdot$$

$$f^{(n_1-2)/2}\left(1+\frac{n_1}{n_2}f\right)^{-(n_1+n_2)/2}\quad(f\geqslant 0)\qquad[3.10]$$

比较方程 3.9 和方程 3.10 可以发现，$t_n^2=F_{1,n}$，而且，随着 n_2 变大，F_{n_1,n_2} 越趋近于 $\chi^2_{n_1}/n_1$，直到 $F_{n,\infty}=\chi^2_n/n$。

对于任意 $n_2>2$，F 的期望值为 $E(F)=n_2/(n_2-2)$，n_2 的取值越大，其越趋近于 1。对于 $n_2>4$，

$$V(F)=\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$$

图 3.10 描述了一些 F 概率密度函数。我们很容易发现，F 分布是正偏的。

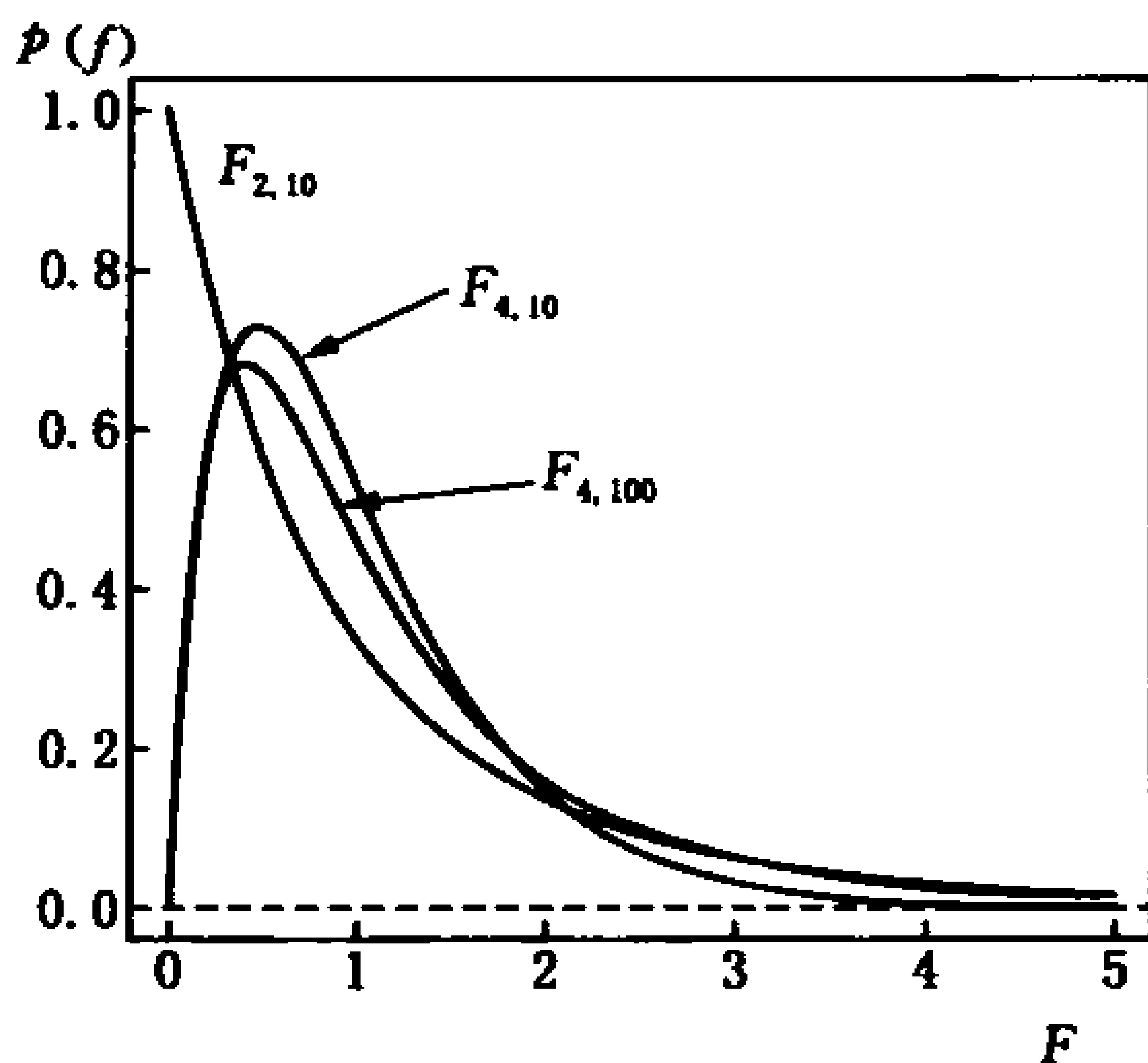


图 3.10 F 密度函数：$F_{2,10}$、$F_{4,10}$ 和 $F_{4,100}$

多元正态分布

一个均值向量为$\boldsymbol{\mu}$、正定方差—协方差矩阵为$\boldsymbol{\Sigma}$的多元正态分布随机向量$\mathbf{x}=[X_1, X_2, \cdots, X_n]'$的联合概率密度可表示为：

$$p(\mathbf{x})=\frac{1}{(2\pi)^{n/2}\sqrt{\det\boldsymbol{\Sigma}}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

通常，我们将其简化为$\mathbf{x}\sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。

如果$\mathbf{x}$是多元正态分布随机向量，那么其包含的元素的边缘分布是单因素正态分布，记做$X_i\sim N(\mu_i, \sigma_i^2)$。[40]给定任意子集的向量，剩下变量的条件分布为$p(\mathbf{x}_1|\mathbf{x}_2)$，其中，$\mathbf{x}=\{\mathbf{x}_1, \mathbf{x}_2\}$也是正态分布的。那么，如果$\mathbf{x}\sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，则有：

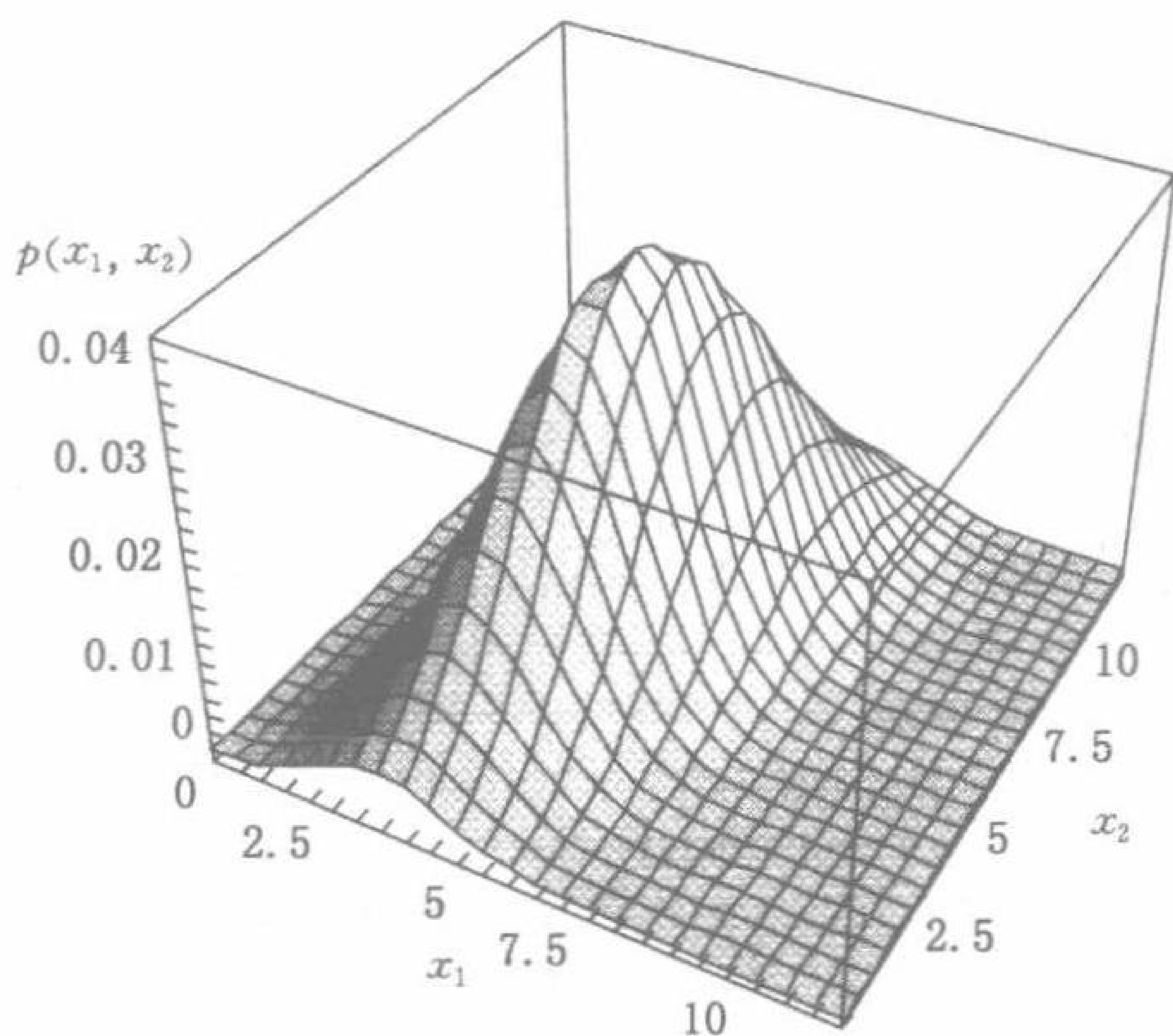
$$\underset{(m\times 1)}{\mathbf{y}}=\underset{(m\times n)}{\mathbf{A}}\underset{(n\times 1)}{\mathbf{x}}$$

秩为$\mathrm{rank}(\mathbf{A})=m\leqslant n$的随机向量$\mathbf{x}$的线性变换，那么，$\mathbf{y}\sim N_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}A')$。如果随机向量$\mathbf{x}$的协方差矩阵$\boldsymbol{\Sigma}$是奇异的，但是$\mathbf{x}$的极大线性无关子集为多元正态分布，那么，我们就说随机向量$\mathbf{x}$遵循奇异正态分布。

有关$\mu_1=5$, $\mu_2=6$, $\sigma_1=1.5$, $\sigma_2=2$, $\rho_{12}=0.5$(如$\sigma_{12}=(0.5)(1.5)(3)=2.25$)的二元正态密度函数请见图3.11。

指数分布

指数分布是一系列以λ为主参数的连续分布，它具有密度函数：



注：其中，密度函数的截面（表示给定其他变量的条件分布）在 x_1 和 x_2 方向上都是正态的。

图 3.11 $\mu_1=5$, $\mu_2=6$, $\sigma_1=1.5$, $\sigma_2=2$, $\rho_{12}=0.5$ 的二元正态密度函数

$$p(x)=\lambda e^{-\lambda x}\quad(x\geqslant 0)$$

X 的期望值和方差分别为 $E(X)=1/\lambda$, $V(X)=1/\lambda^2$。图 3.12 描述了几个具有不同参数的指数分布。指数分布具有高度的正偏性，因此，当事件出现的“风险”在观测期中是一个常数时，它常被用于时间到事件数据的建模。

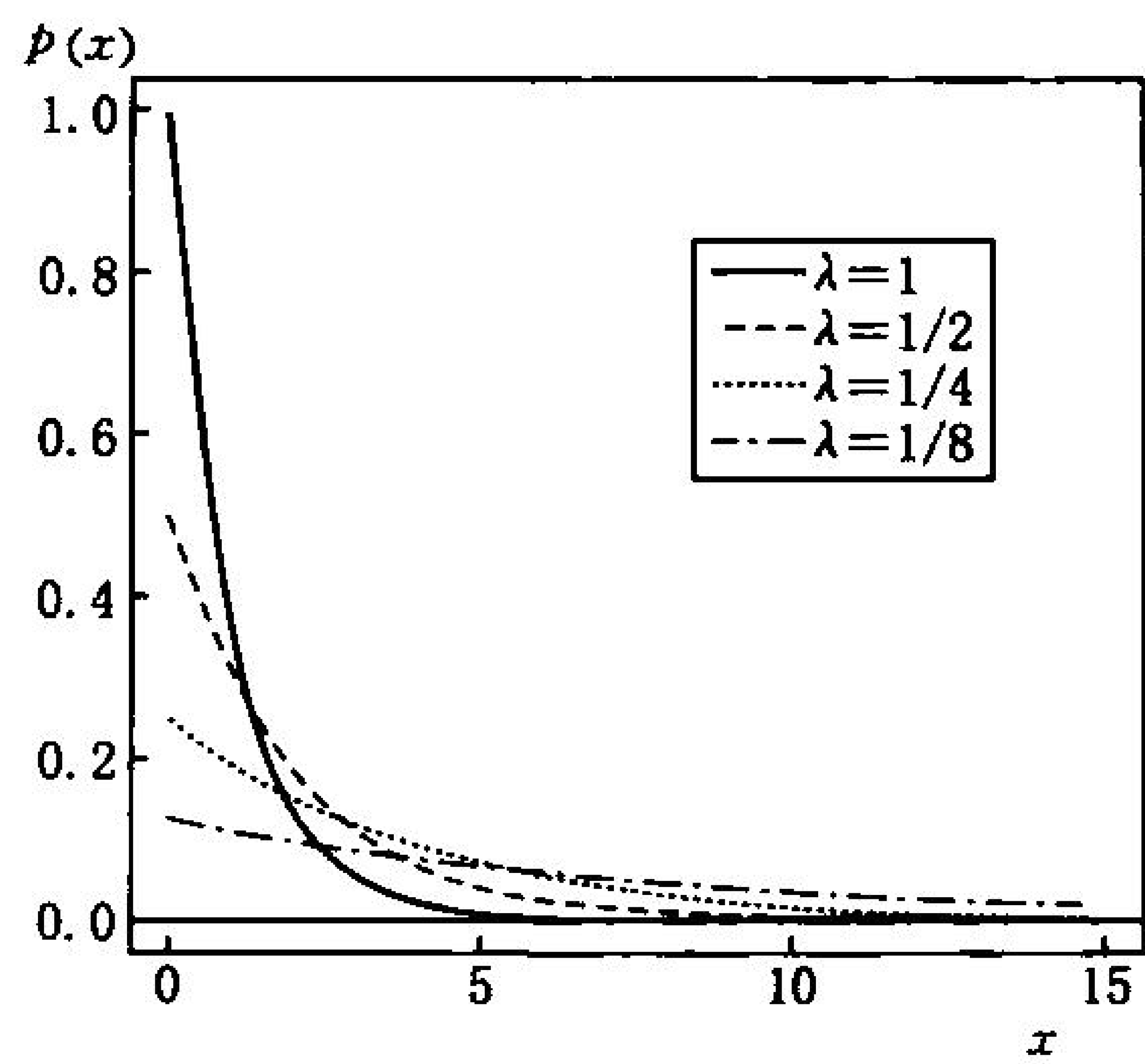


图 3.12 关于不同参数 λ 的指数分布

逆高斯分布

逆高斯分布是关于两个系数 μ 和 λ 的连续分布，它具有密度函数：

$$p(x)=\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left[-\frac{\lambda(x-\mu)^2}{2x\mu^2}\right]\quad(x>0)$$

X 的期望值和方差分别为 $E(X)=\mu$, $V(X)=\mu^3/\lambda$。图 3.13 描绘了几个逆高斯分布。逆高斯分布的方差随着其均值的增大而增大；偏度随着 μ 的增大而增大，随着 λ 的增大而减小。

逆高斯矩阵和伽马分布（下面即将介绍）常用来对非负连续数据建模。

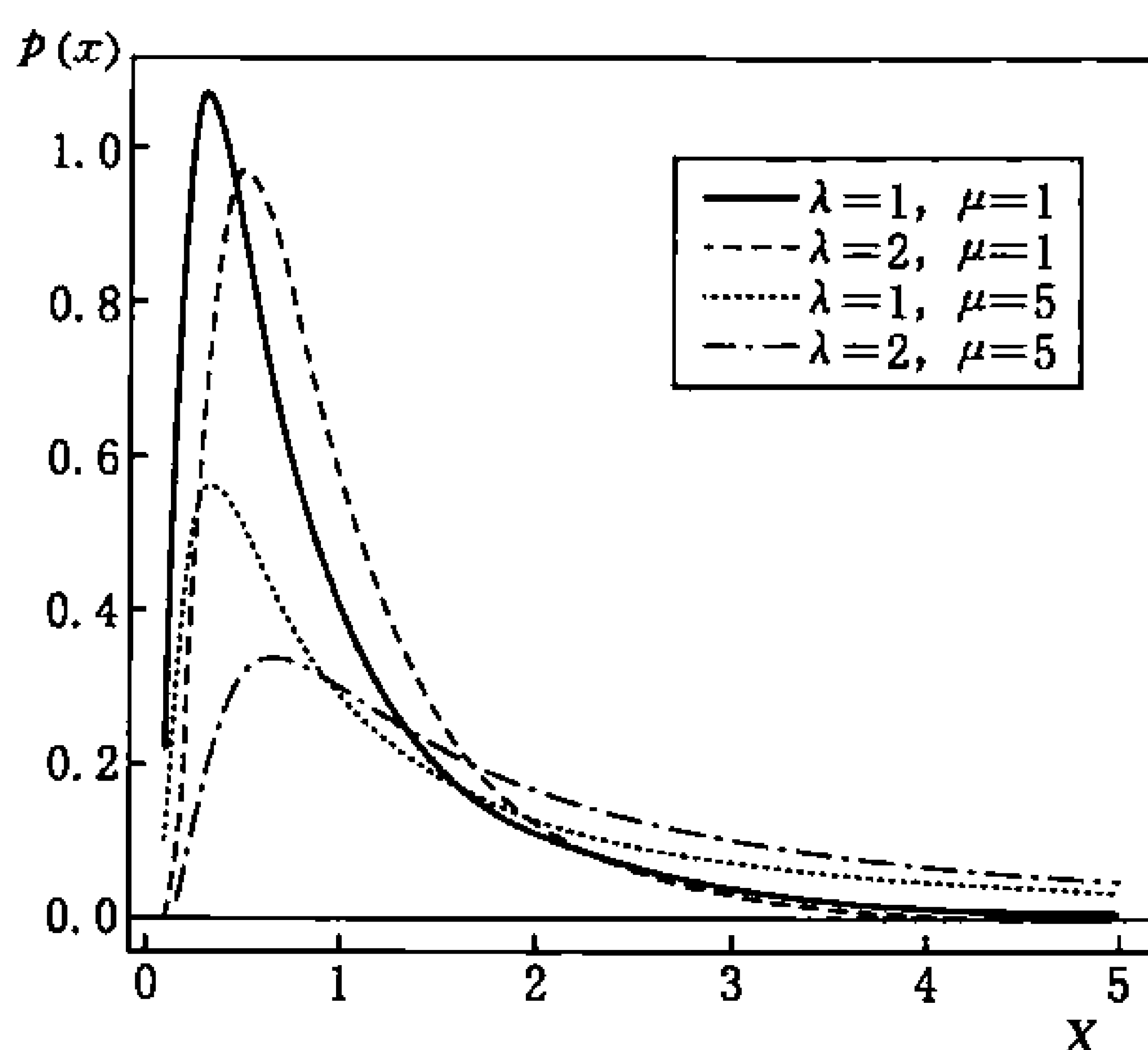


图 3.13 关于不同参数 λ 和 μ 的逆高斯分布

伽马分布

伽马分布属于连续分布，它是具有尺度参数 $\omega>0$ 和形

状参数 $\Psi > 0$ 的概率密度函数：

$$p(x) = \left(\frac{x}{\omega}\right)^{\Psi-1} \frac{\exp\left(\frac{-x}{\omega}\right)}{\omega\Gamma(\Psi)} \quad (x > 0)$$

其中，$\Gamma(\cdot)$为伽马函数(见方程 3.8)。伽马分布的期望值和方差分别为 $E(X) = \omega\Psi$, $V(X) = \omega\Psi^2$。图 3.14 描述了在尺度 $\omega = 1$ 下，不同形状参数 Ψ 的伽马分布(改变尺度参数仅会使图像在水平轴上平移)。

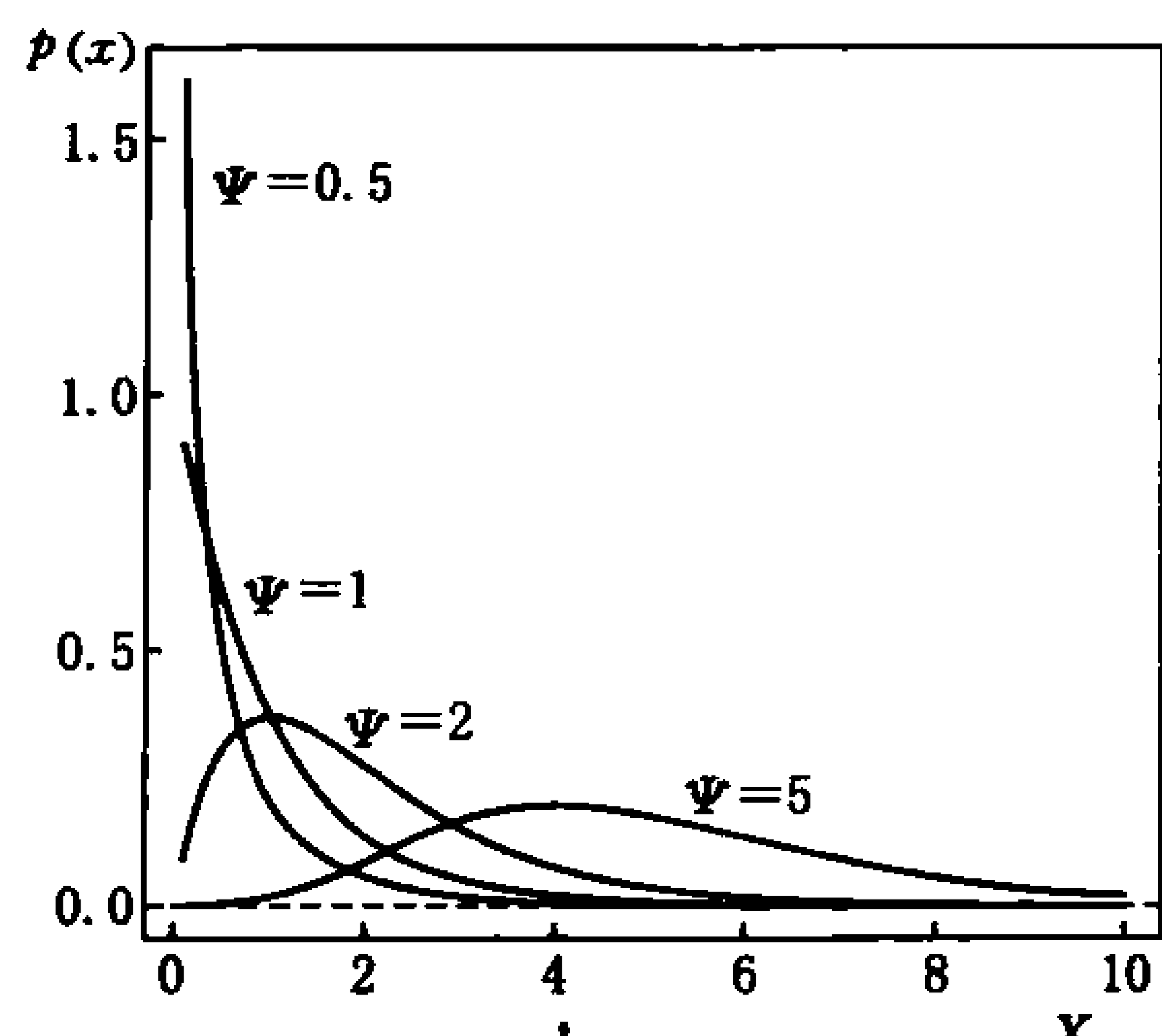


图 3.14　不同形状参数 Ψ 的伽马分布($\omega = 1$)

如果 $X_1, X_2, \cdots, X_k$ 是具有相同尺度参数ω、不同形状参数 $\Psi_1, \Psi_2, \cdots, \Psi_k$ 的独立伽马随机变量，那么 $X = X_1 + X_2 + \cdots + X_k$ 为具有尺度参数ω 和形状参数 $\Psi = \Psi_1 + \Psi_2 + \cdots + \Psi_k$ 的伽马分布。

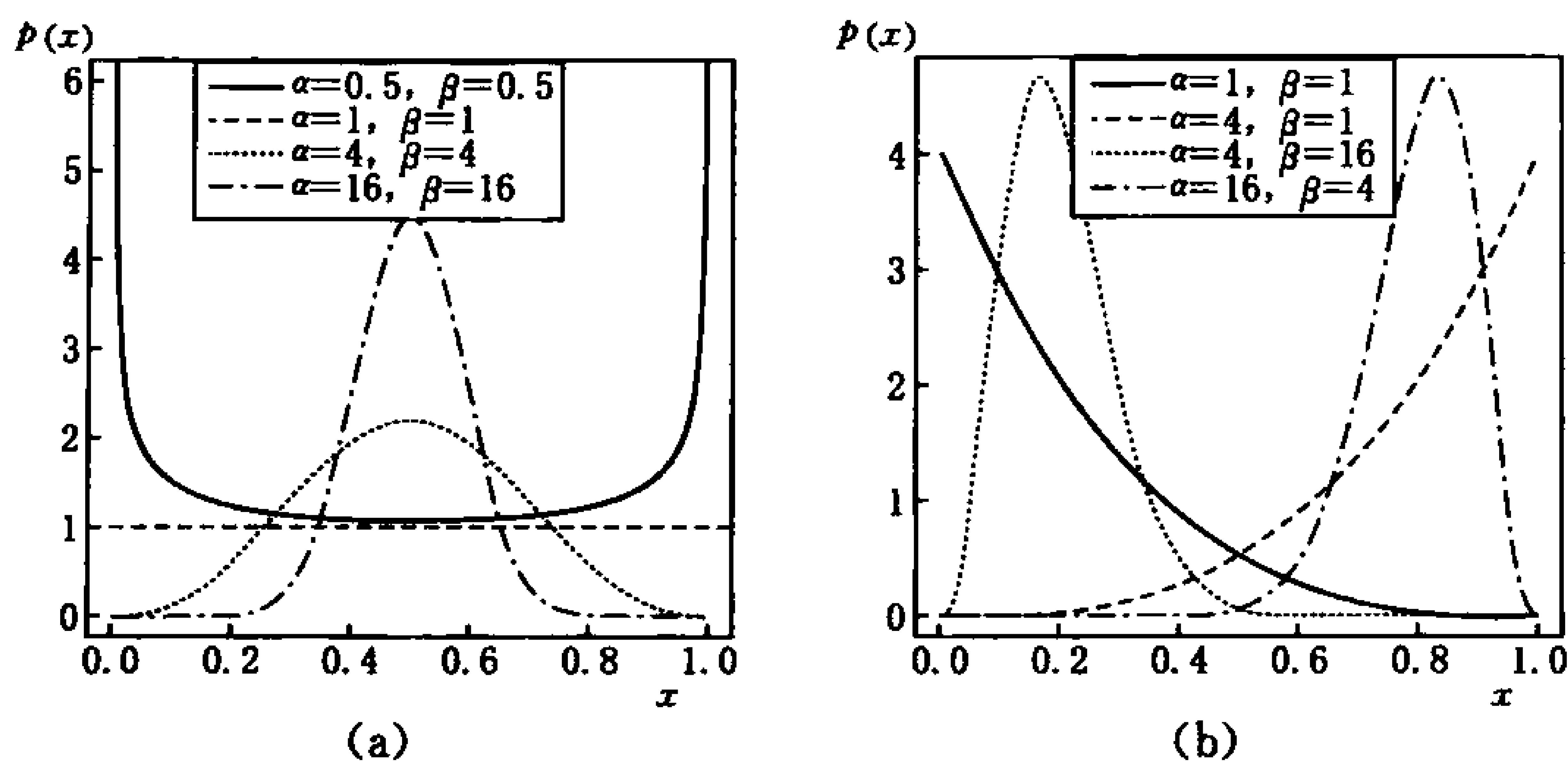
含有 n 个自由度的卡方分布和具有尺度参数$\omega = 2$ 及形状参数 $\Psi = n/2$ 的伽马分布是相等的。主参数为λ 的指数分布和具有尺度参数 $\omega = 1/\lambda$ 及形状参数 $\Psi = 1$ 的伽马分布是相等的。

贝塔分布

贝塔分布是包含两个形状参数 $\alpha > 0$，$\beta > 0$ 的连续分布，它具有密度函数：

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (0 \leqslant x \leqslant 1)$$

其中，$B(\alpha, \beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ 为贝塔函数。贝塔分布的期望值和方差分别为 $E(X) = \alpha/(\alpha+\beta)$，$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$。因此，期望值取决于参数 α、β 的相对大小。如当 $\alpha = \beta$ 时，$E(X) = 0.5$。偏态也同样取决于参数的相对值，且当 $\alpha = \beta$ 时，分布是对称的。方差随 α、β 的增大而减小。图 3.15 描述了几个贝塔分布。很明显，贝塔分布的变化很灵活。



注：在图 3.15(a)中，很明显，当 $\alpha = \beta = 1$ 时，贝塔分布退化为矩分布。图 3.15(a)为对称的贝塔分布，图 3.15(b)为反对称的贝塔分布。

图 3.15　不同 α、β 组合的贝塔分布

第 4 节 | 渐近分布理论:初步介绍

有时,因为很难确定统计估计量的小样本性质,所以研究一个估计量随着样本增大的表现就变得尤为重要。渐近分布理论就为这类研究提供了工具。在本章节,我仅对该理论进行概述,更完整的叙述可参考其他相关书籍。

极限概率

渐近分布理论常被应用于随机变量序列中。但是,我们有必要先考虑非随机无限序列$\{a_1, a_2, \cdots, a_n, \cdots\}$。关于“非随机”,我指的是每一个$a_n$而非随机变量是固定的。读者可能会注意到,如果对于任意无限小的数ε,总是存在一个正数$n(\varepsilon)$,对于所有的$n>n(\varepsilon)$,有$|a_n-a|<\varepsilon$,那么我们称该数列存在极限a。换句话说,只要n足够大,a_n就可以任意地接近a。$n(\varepsilon)$强调了n值取决于我们所选择的标准ε(请参见前文有关函数极限的定义)。为了使表述更简洁,我们可以用表达式$\lim\limits_{n\to\infty} a_n = a$。例如,若$a_n = 1+1/n$,那么$\lim\limits_{n\to\infty} a_n = 1$。图 3.16 描述了该数列及其极限。

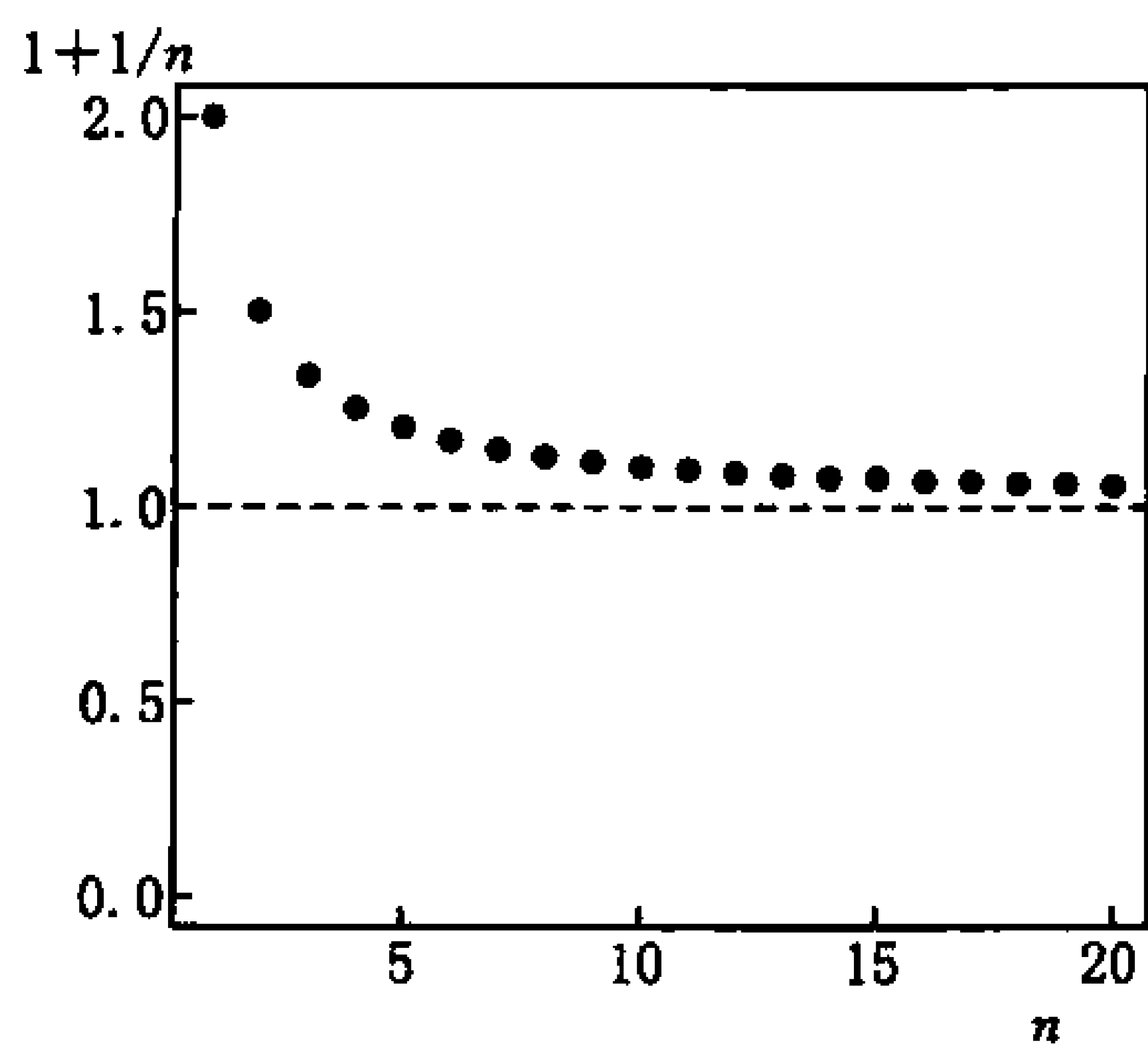
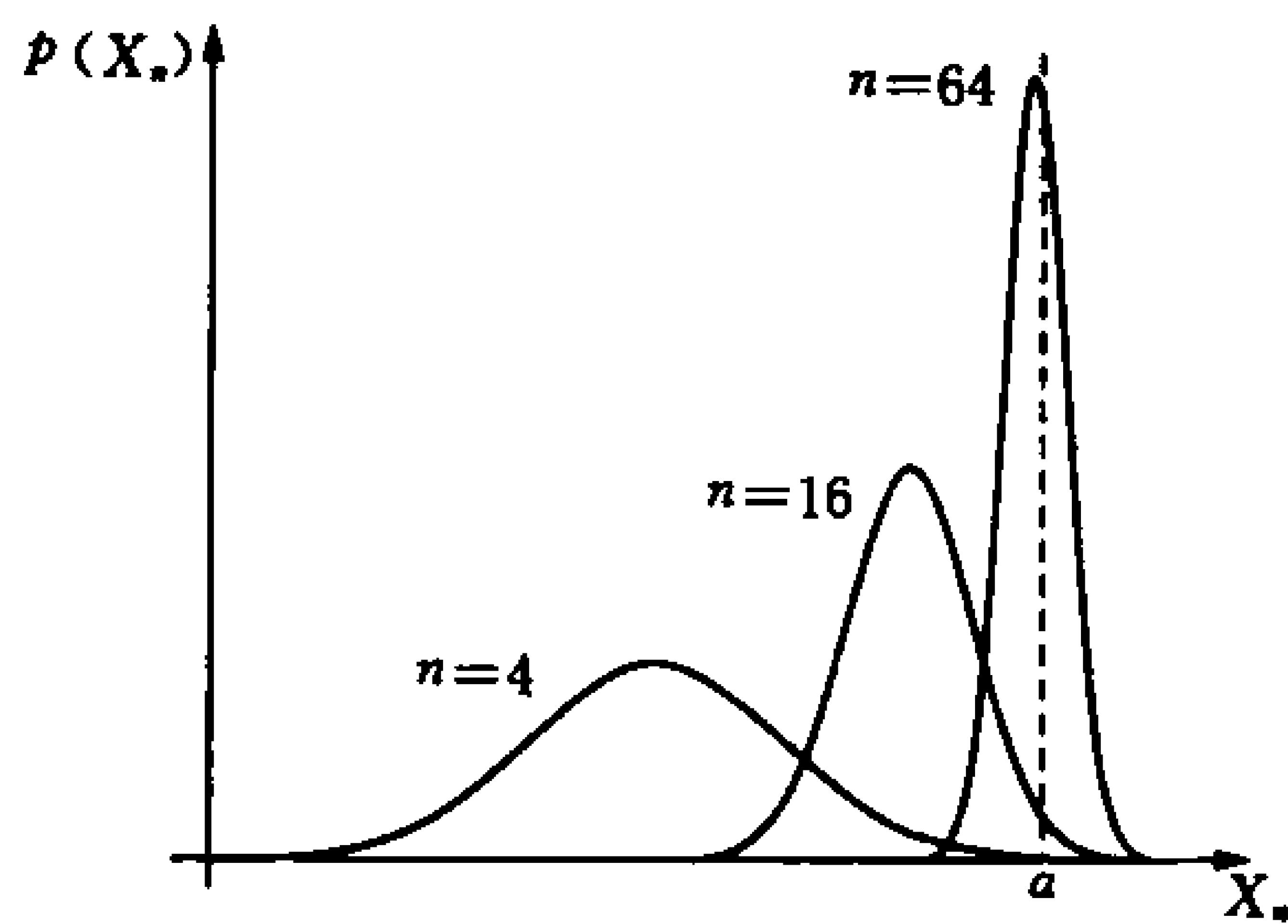


图 3.16 具有极限 $a=1$ 的序列 $a_n=1+1/n$ 的前20个值

我们现在考虑随机变量序列$\{X_1, X_2, \cdots, X_n, \cdots\}$。在应用统计中,$X$ 为估计量,n 为该估计量的样本大小。让 $p_n \equiv \Pr(|X_n-a|<\delta)$,其中 a 是一个常数,δ 是一个很小的整数。我们可以把 p_n 想象成 X_n 逐渐接近 a 的概率。假设非随机概率序列$\{p_1, p_2, \cdots, p_n, \cdots\}$以 1 为极限[41],即 $\lim_{n\to\infty}\Pr(|X_n-a|<\delta)=1$。那么随着 n 的增大,随机变量 X_n 将在 a 的小范围区域内更接近 a,图 3.17 描述了此情形。如果无论 δ 多么小,该结果都成立,那么我们说 a 是 X_n 的概率极限,表示为 $\text{plim}\, X_n=a$。为方便起见,我们可以把 n 去掉,记作 $\text{plim}\, X=a$。



注:随着 n 的增大,X_n 将越来越接近 a。

图 3.17 $\text{plim}\, X_n=a$

概率极限具有如下重要性质：假如 $\text{plim}\, X = a$，且 $Y = f(X)$ 为 X 的连续函数，那么，$\text{plim}\, Y = f(a)$。同样，如果 $\text{plim}\, X = a$，$\text{plim}\, Y = b$，$Z = f(X, Y)$ 为 X 和 Y 的连续函数，那么 $Z = f(a, b)$。

渐近期望均值和方差

回到随机变量序列$\{X_1, X_2, \cdots, X_n, \cdots\}$，并令 μ_n 为 X_n 的期望值。那么，$\{\mu_1, \mu_2, \cdots, \mu_n, \cdots\}$为一个非随机序列。如果该序列趋近于一个极限 μ，那么我们说 μ 为 X 的渐近期望值，记做 $\epsilon(X)$。

尽管我们会很自然地把渐近方差的定义序列与方差的极限进行类比，但是该定义无法让人满意，因为在许多情况下（下面将举例说明），$\lim_{n\to\infty} V(X_n) = 0$。假设我们计算一个从均值为 μ、方差为 σ^2 的总体中抽取的（大小为 n）样本均值，将其记做 $\bar{X}_n$。由初等统计学可知，$E(\bar{X}_n) = \mu$，另外，

$$V(\bar{X}_n) = E[(\bar{X}_n - \mu)^2] = \frac{\sigma^2}{n}$$

因此，$\lim_{n\to\infty} V(\bar{X}_n) = 0$，把$\sqrt{n}$加入中括号内，有 $E\{[\sqrt{n}(\bar{X}_n - \mu)]^2\} = \sigma^2$，将其除以 n 然后取极限即可得到我们想要的结果，此时样本均值的渐近方差为：

$$\begin{aligned} \upsilon(\bar{X}) &\equiv \lim_{n\to\infty} \frac{1}{n} E\{[\sqrt{n}(\bar{X}_n - \mu)]^2\} \\ &= \frac{1}{n}\epsilon\{[\sqrt{n}(\bar{X}_n - \mu)]^2\} \\ &= \frac{\sigma^2}{n} \end{aligned}$$

该结果没有什么特别之处,因为 $V(\bar{X}) = \upsilon(\bar{X})$。事实上,这与一开始给出的渐近方差的定义是等价的。在实际应用中,当有限样本方差不可求时,还是有可能找到渐近方差的。此时,我们可以将渐近结果当做大样本的近似。

通常,如果 X_n 的期望值为 μ_n,那么,X 的渐近方差定义为[42]:

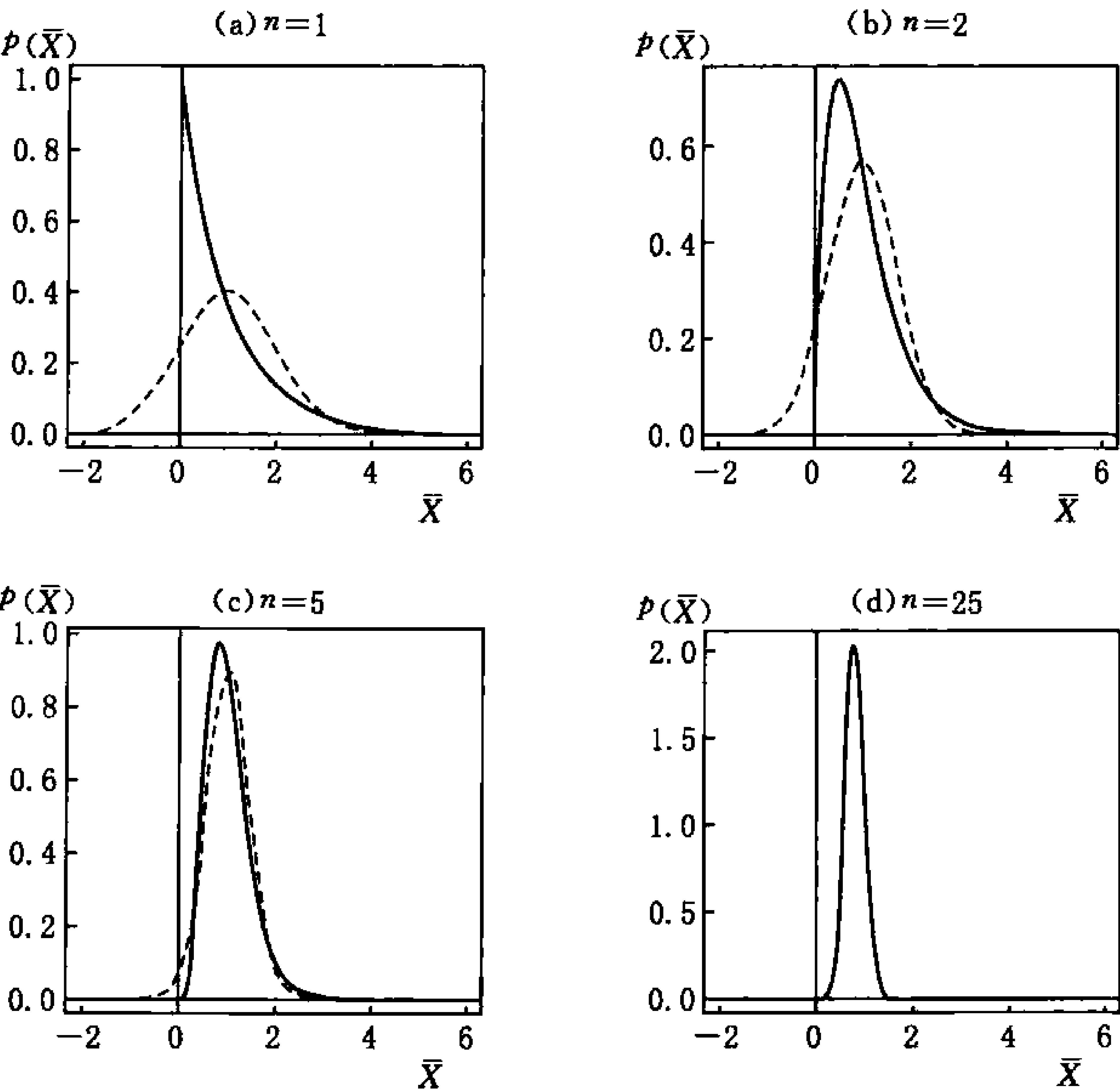
$$\upsilon(X) \equiv \frac{1}{n}\varepsilon\{[\sqrt{n}(X_n - \mu_n)]^2\} \qquad [3.11]$$

渐近分布

让$\{P_1, P_2, \cdots, P_n, \cdots\}$代表随机变量序列$\{X_1, X_2, \cdots, X_n, \cdots\}$的累积分布函数。假如对于随机变量的所有值 x 和任意无论多小的数 ε,我们总能找到一个足够大的数 $n(\varepsilon)$,那么,对于所有的 $n > n(\varepsilon)$,都有 $|P_n(x) - P(x)| < \varepsilon$;那么,我们说 X 的累积分布函数收敛于渐近分布 P。

中心极限定理描述了当一组独立同分布,且具有有限的期望值和方差的随机变量的均值遵循近似正态分布时,该近似过程会随着随机变量数目的增加而加强。例如,有一个样本大小为 n、主参数为 λ,且高度偏斜的指数分布,其均值 μ 和方差 σ^2 都为 1。我们知道,指数分布是伽马分布的一个特例,其形状参数 $\Psi = 1$,尺度参数 $\omega = 1/\lambda$,那么,样本的和 $\sum_{i=1}^{n} X_i$(即 $n\bar{X}$)是形状参数 $\Psi = n$、尺度参数 $\omega = 1$ 的伽马分布。图 3.18 描述了从指数分布总体得到的样本均值 $\bar{X}$ 的抽样分布密度函数随样本量大小的变化,且每一种情况都比

较了 $\bar{X}$ 的真实伽马样本分布与近似正态分布 $N(1, 1/n)$，正态近似随样本量的增加而越发精确（而 $\bar{X}$ 的抽样分布的方差随之减小）。



注：图 3.18(a)，$n = 1$ 所对应的 X 的总体分布。在每一张图中，实线为真实（伽马）抽样分布 $\bar{X}$ 的密度函数，虚线为正态近似 $N(1, 1/n)$ 的密度函数。

图 3.18　中心极限定理：从指数分布总体（主参数为 $\lambda = 1$）得到的（样本量 n 的大小不同）样本均值 $\bar{X}$ 的抽样分布

随机向量与随机矩阵

我们将以上结果扩展到向量和矩阵中，得到：当 $\text{plim}\, X_i = a_i (i = 1, 2, \cdots, m)$ 时，$\text{plim} \underset{(m\times1)}{\mathbf{x}} = \underset{(m\times1)}{\mathbf{a}}$。$\text{plim} \underset{(m\times p)}{\mathbf{X}} = \underset{(m\times p)}{\mathbf{A}}$ 意

味着，对于所有的 i 和 j，$\text{plim}\, X_{ij} = a_{ij}$。随机向量 $\underset{(m\times 1)}{\mathbf{x}}$ 的近似期望值定义为由其中元素的近似期望值组成的向量，即 $\boldsymbol{\mu} = \varepsilon(\mathbf{x}) \equiv [\varepsilon(X_1), \varepsilon(X_2), \cdots, \varepsilon(X_m)]'$。$\mathbf{x}$ 的渐近方差—协方差矩阵定义为：

$$\upsilon(\mathbf{x}) \equiv \frac{1}{n}\varepsilon\{[\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}_n)][\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}_n)]'\}$$

第 5 节 | 统计估计量的属性[43]

一个样本统计量（即一个有关样本中众多观测的函数）的估计量是用来估计总体参数的。由于其数值因样本不同而异，因此，估计量是一个随机变量。估计是特定样本估计量的数值。估计量的概率分布称为“抽样分布”，该分布所对应的方差称为估计量的“抽样方差”。

偏差

如果 $E(A)=\alpha$，那么我们说参数 α 的估计量 A 是无偏的。因此，$E(A)-\alpha$ 即 A 的偏差。

假设我们从均值为 μ、方差为 σ^2 的总体中得到 n 个观测 X_i，那么，我们说样本均值 $\bar{X}\equiv\sum X_i/n$ 是 μ 的无偏估计量，同时，

$$S_*^2\equiv\frac{\sum(X_i-\bar{X})^2}{n} \qquad [3.12]$$

S_*^2 是 σ^2 的有偏估计量，因为 $E(S_*^2)=[(n-1)/n]\sigma^2$，$S_*^2$ 的偏差因此等于 $-\sigma^2/n$。有关抽样分布的无偏及有偏估计量，请见图 3.19。

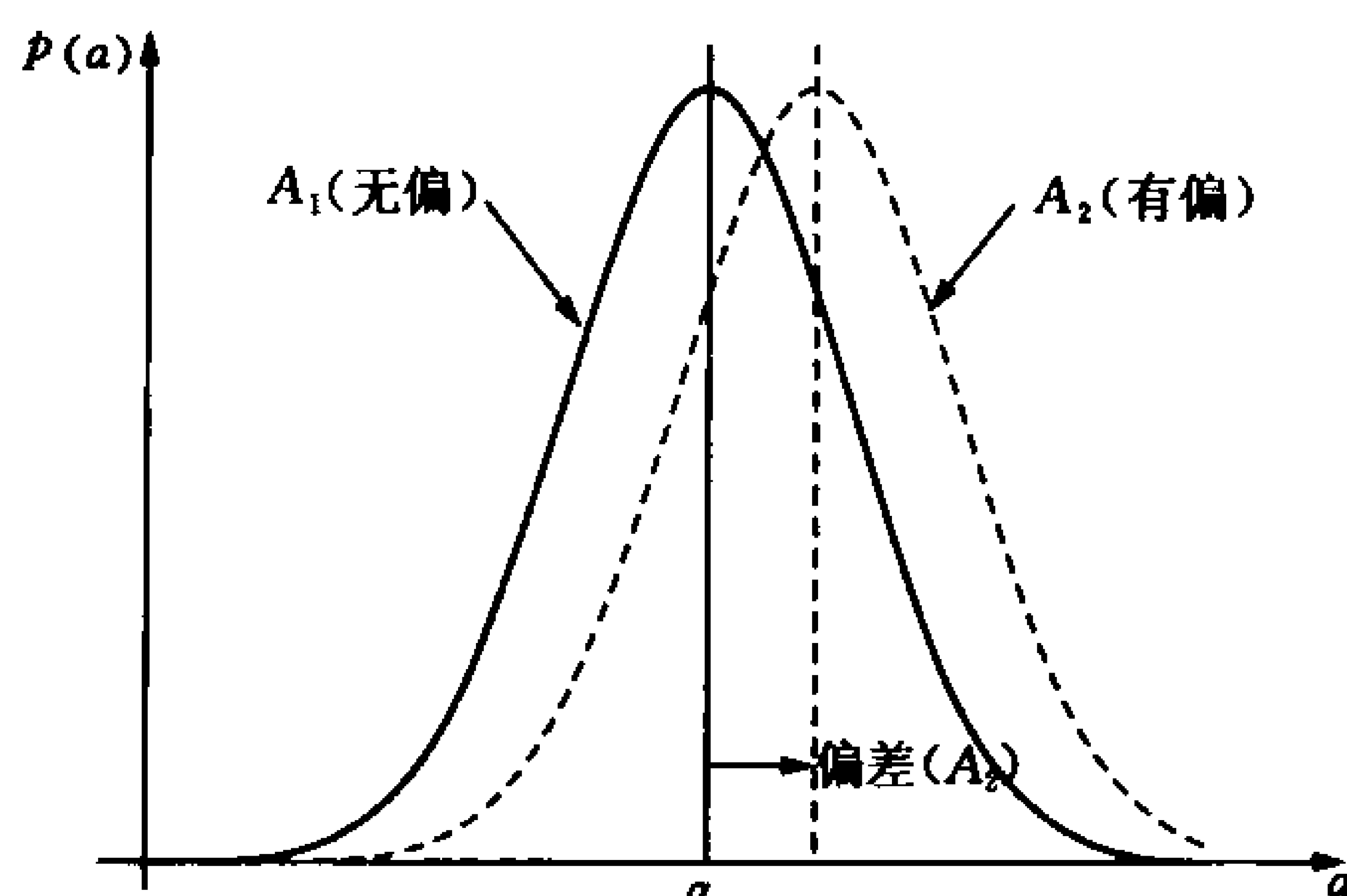


图 3.19 因为 $E(A_1)=\alpha$,所以估计量 A_1 是 α 的无偏估计量;因为 $E(A_2)>\alpha$,因此估计量 A_2 是正偏的

渐近偏差

参数 α 的估计量 A 的渐近偏差是 $\varepsilon(A)-\alpha$,那么,如果 $\varepsilon(A)=a$,则估计量 A 是无偏的。由于当 $n\to\infty$ 时,$-\sigma^2/n\to 0$,因此,S_*^2 是渐近无偏的。

均方误差与有效性

一个估计量是无偏的意味着,其重复样本的平均数值和总体估计参数相同。很明显,该特征应该是估计量最理想的性质。但是,如果样本估计量和总体估计参数不接近的话,那么,该估计量则是无用的。对于期望值,一些样本的较大的负估计误差可以抵消其他样本的较大的正估计误差。

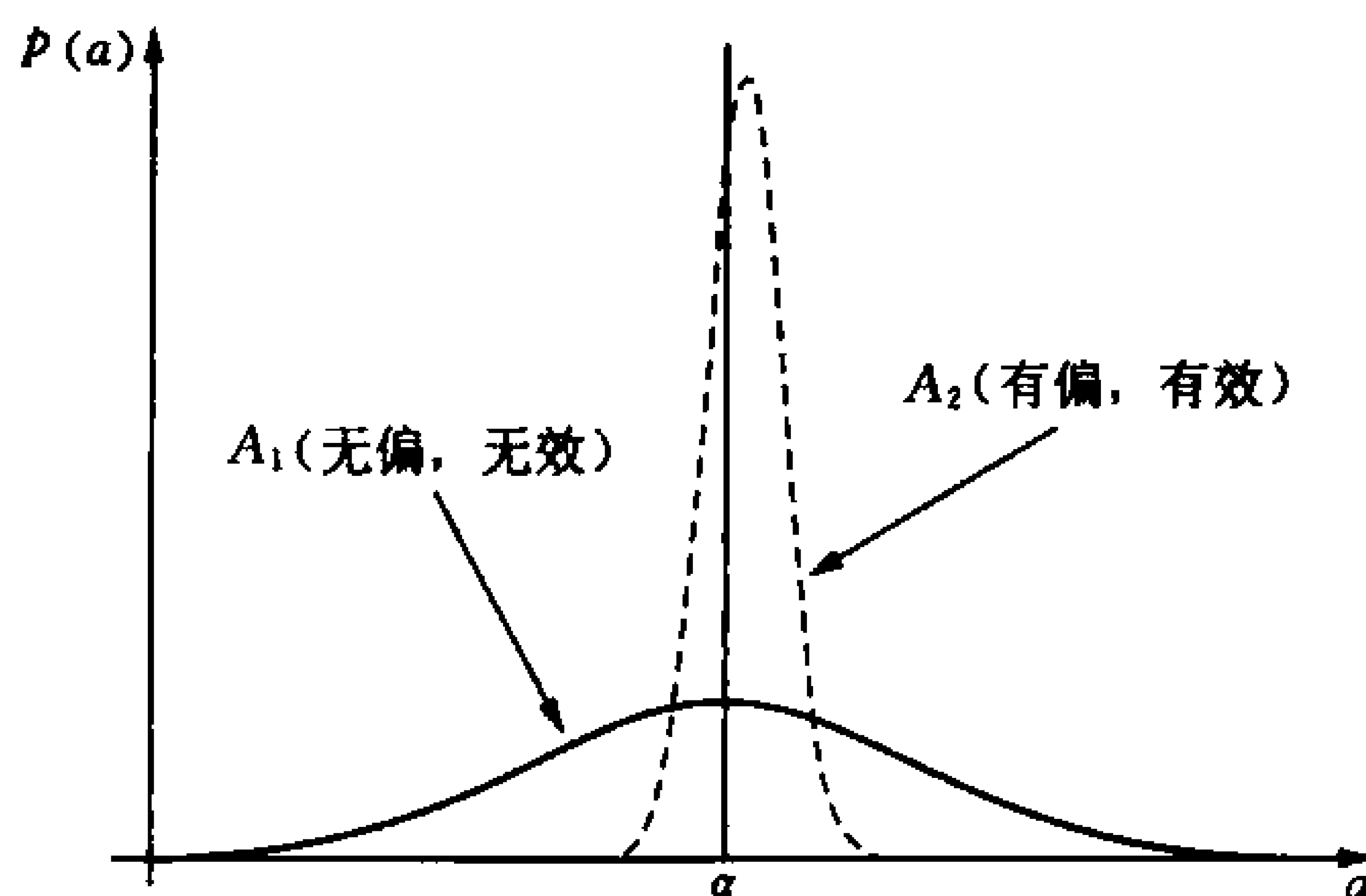
参数 α 的估计量的均方误差(MSE)是估计量与参数之间的差异平方的均值,即 $MSE(A)\equiv E[(A-\alpha)^2]$。估计量的有效性与其均方误差成反比。通常,我们比较倾向于有效的估计量。

由于 $E(A)=\alpha$,因此,一个无偏估计量的均方误差,简

单地说，就是其抽样方差。而对于一个有偏估计量，

$$\begin{aligned} MES(A) &= E[(A-\alpha)^2] = E\{[A-E(A)+E(A)-\alpha]^2\} \\ &= E\{[A-E(A)]^2\}+[E(A)-\alpha]^2+2[E(A) \\ &\quad -E(A)][E(A)-\alpha] \\ &= V(A)+[bias(A)]^2+0 \end{aligned}$$

当一个估计量的有效性增加时，其抽样方差及偏差会减小。那么，比较两个估计量抽样方差上的优势可以更多地补偿其偏差劣势，如图 3.20 所示。



注：尽管估计量 A_2 是有偏的，但是 A_2 相对于无偏估计量 A_1，是参数 α 的一个更有效的估计量，其原因在于，A_2 的小方差性可以部分补偿其偏差性。

图 3.20 估计量的相对有效性

渐近有效性

渐近有效性与渐近均方误差（*AMSE*）成反比，且渐近均方误差是渐近方差和渐近偏差平方的和。

一致性

如果 $\text{plim} A = \alpha$，那么参数 α 的估计量 A 是一致的。一

致性的充分(非必要)条件是估计量本身是渐近无偏的,且抽样方差随着 n 的增加趋近于 0。该条件暗示了估计量的均方误差的极限为 0。图 3.17 描述了 a 的估计量 X 的一致性。方程 3.12 表示,估计量 S_*^2 是总体方差 σ^2 的一致估计量,尽管在有限样本中,它本身存在偏差。

充分性

充分性的概念比无偏、有效及一致更抽象:如果在样本中,统计量详尽地表达了参数 α 的所有信息,那么,基于观测值的统计量 S 符合充分性条件,或者可以说,假设观测值 $X_1, X_2, \cdots, X_n$ 是从以 α 为参数的概率分布中得来的。我们让统计量 $S = f(X_1, X_2, \cdots, X_n)$,如果观测值的概率分布是以 S 的数值为条件的,也就是说 $p(x_1, x_2, \cdots, x_n \mid S = s)$ 与 α 无关,那么 S 即一个 α 的充分统计量。注意,充分统计量 S 不是参数 α 的估计量。

要描述充分性,我们可以假设 n 个观测都是独立采样得来的,对于每个观测,X_i 为 1 的概率为 π,为 0 的概率为 $1-\pi$。即,X_i 是独立同分布的伯努利随机变量。在这里,我会证明样本总和 $S \equiv \sum_{i=1}^{n} X_i$ 是 π 的充分统计量。如果我们已知 S 的值 s,则对于 S 的不同种组合 ($s = 0$; $s = 1$),数目为 $\binom{n}{s}$,且每种组合的可能性为 $1\Big/\binom{n}{s}$。我们知道,随机变量 S 遵循一个二项分布,由于其概率与 π 无关,因此,统计量 S 是 π 的充分统计量。同理,样本比例 $P \equiv S/n$ 也是一个充分统计量。样本比例 P(而不是总和 S)是 π 的估计量。

充分性的概率可以延伸到一组参数和统计量上：已知一个样本（可能为多元）中的观测为 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$，如果观测的条件分布决定了 $\mathbf{s}$ 与 $\boldsymbol{\alpha}$ 无关，那么，向量统计量 $s = [S_1, S_2, \cdots, S_p]' \equiv f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ 是参数 $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_k]'$ 的联合充分统计量。例如，独立随机变量的均值 $\bar{X}$ 和 S^2 分别是正态分布参数 μ 和 σ^2 的联合充分统计量（因为样本总和 $\sum X_i$ 与平方和 $\sum X_i^2$ 所包含的信息与 $\bar{X}$ 和 S^2 相同）。如果没有更小的充分统计量组，那么，该组充分统计量则是最低充分统计量组。

稳健性

当一个估计量的有效性（及其相对其他估计量的有效性）不极大地依赖于数据分布时，那么我们说该估计量是稳健的。

还有另一种稳健性，称为“效度稳健性”，我们要将它与有效稳健性相区分。对于统计推理过程，如果其效度不极大地依赖于数据分布，那么，我们说它是稳健的。因此，即使检验违反了分布假设（如正态分布假设），稳健性假设检验的 p 值仍可看做是近似准确的。同样，如果置信区间的覆盖率与之前所陈述的相同（例如，一个 95％置信区间覆盖了差不多 95％的样本），即使有时会违反分布假设，但我们仍说该置信区间是稳健的。当一个检验或者置信区间是基于一个无效估计量时，如果检验的统计功效很低，或者置信区间很宽，那么，该检验或者置信区间的效度稳健性就很低。

要具体描述有效稳健性，我们就要把重点放在估计一个

对称分布的中心 μ 上。[44]只要 X 存在有限方差 σ^2，那么，样本均值 $\bar{X}$ 的方差为 $V(X) = \sigma^2/n$，这里的 n 是样本量（与基本统计的结果一致），且样本中位数的方差与 X 的分布有关：

$$V(\text{median}) \approx \frac{1}{4n[p(x_{0.5})]^2}$$

其中，$p(x_{0.5})$ 为 X 为总体中位数时的密度。

运用到正态分布的总体上则有 $X \sim N(\mu, \sigma^2)$，中位数方差为 $V(\text{median}) = \pi\sigma^2/2n$，因此，样本均值相对于中位数是一个相对有效的统计量：

$$\frac{V(\text{median})}{V(\bar{X})} = \frac{\pi\sigma^2/2n}{\sigma^2/n} = \frac{\pi}{2} \approx 1.57$$

为了保证准确性，用样本中位数来估计 μ 所用的样本量是用样本均值进行估计时的 1.57 倍。

相反，假设 X 服从自由度为 3 的 t 分布，该分布相对于正态分布，尾部较重且较长。那么，$\sigma^2 = 3/(3-2) = 3$，$p_{(x_{0.5})} = p(0) = 0.3675$，因此，

$$V(\bar{X}) = \frac{3}{n}$$

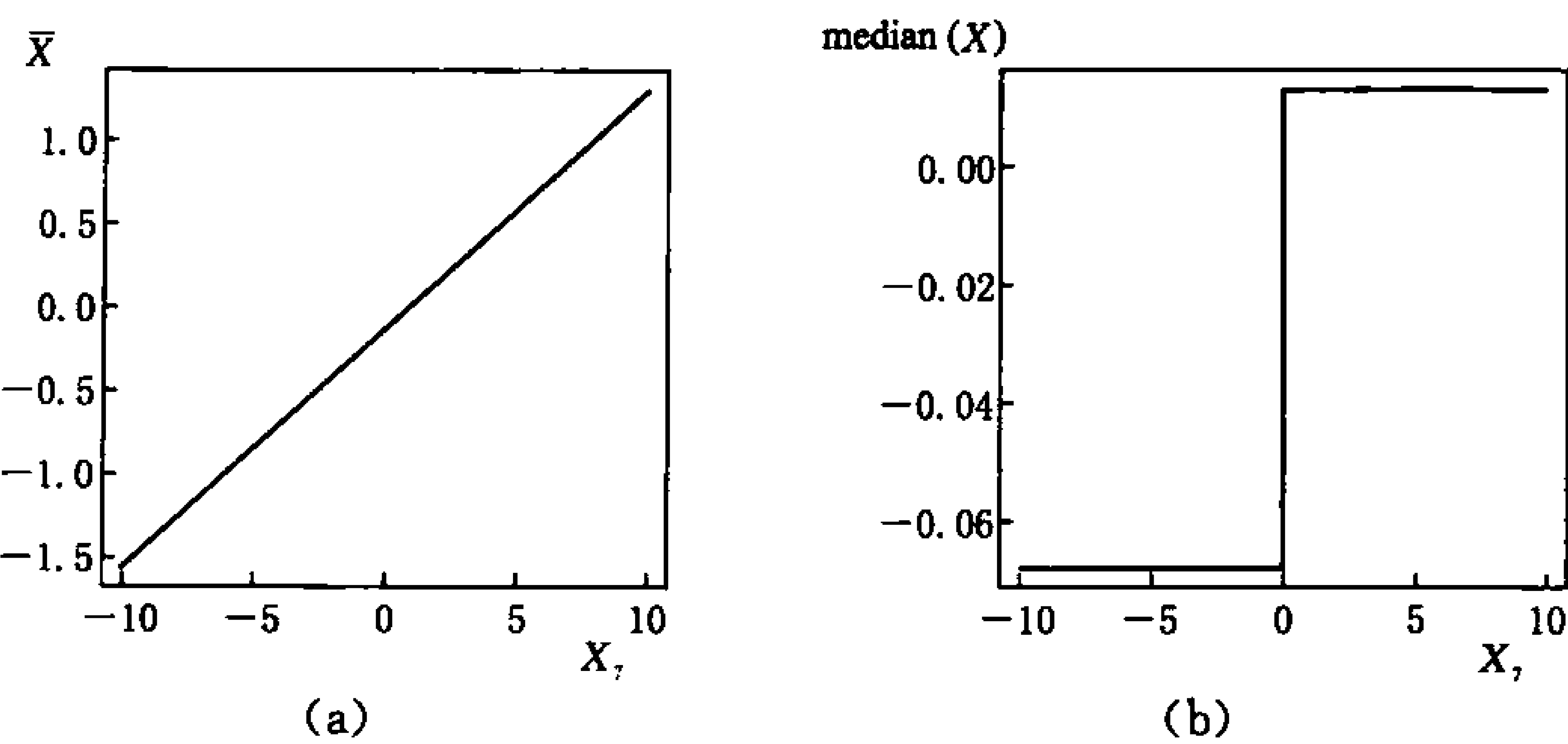
$$V(\text{median}) = \frac{1}{4n(0.3675^2)} = \frac{1.851}{n}$$

对于此例，均值只有 $1.851/3 = 0.617(62\%)$。因此，在这里，均值与中位数一样有效。

稳健性对于异常数据具有耐抗性，一个耐抗估计量不会被小部分的异常数据所影响。均值对异常值的耐抗性比较小，这点很容易证明：从一个标准正态分布中抽取一个含有六个观测的样本，如下所示：

$$X_1 = -0.068 \quad X_2 = -1.282 \quad X_3 = 0.013$$
$$X_4 = 0.141 \quad X_5 = -0.980 \quad X_6 = 1.263 \quad [3.13]$$

这些值的均值为 $\bar{X} = -0.152$。如果我们想加入第七个观测，即 X_7，它可取从 -10 到 $+10$（或者范围更广一些，例如，从 $-\infty$ 到 $+\infty$）的一切可能数值。该结果称为均值的“影响函数”，如图 3.21(a)所示。很明显，随着 X_7 的取值趋向于极值，样本均值也不断增大。



注：中位数对函数的影响是有界的，但均值却不是。注意，两张图纵坐标的刻度不同。

图 3.21 样本均值(a)和中位数(b)的影响函数 $X_1 = -0.068,\ X_2 = -1.282,\ X_3 = 0.013,\ X_4 = 0.141,\ X_5 = -0.980,\ X_6 = 1.263$

与估计耐抗性相关的一个概念称为估计值的“崩溃点”。崩溃点是估计值可以耐受而不会被任意异常大的值所影响的“坏”数据部分。均值的崩溃点是 0，因为正如我们所看到的，一个不好的观测可以任意地改变均值的大小。相反，中位数的崩溃点为 50%，原因在于，即使有一半的数据是“坏”的，中位数也不会被完全影响。

M 估计

用均值将最小二乘目标函数最小化后得到：

$$\sum_{i=1}^{n}\rho_{LS}(X_i-\hat{\mu})\equiv\sum_{i=1}^{n}(X_i-\hat{\mu})^2$$

该均值影响函数的形状为目标函数对残差 $E\equiv X-\hat{\mu}$ 求导的结果：

$$\psi_{LS}(E)\equiv\rho'_{LS}(E)=2E$$

影响函数因此和 E 成正比。那么，将最小二乘目标函数重新定义为 $\rho_{LS}(E)=\frac{1}{2}E^2$ 会更加方便，这样的话，$\psi_{LS}(E)=E$。

现在考虑样本中位数是 μ 的估计值的情况。中位数最小化了最小绝对值(LAV)的目标函数：

$$\sum_{i=1}^{n}\rho_{LAV}(E_i)\equiv\sum_{i=1}^{n}\rho_{LAV}(X_i-\hat{\mu})\equiv\sum_{i=1}^{n}|X_i-\hat{\mu}|$$

结果我们发现，中位数对异常值的耐抗性比均值强得多。有关中位数的影响函数请参见图 3.21(b)。与均值相反，中位数在观测差异上的影响是有界的。之前提到过，目标函数的导数决定了影响函数的形状[45]：

$$\psi_{LAV}(E)\equiv\rho'_{LAV}=\begin{cases}1 & (E>0)\\0 & (E=0)\\-1 & (E<0)\end{cases}$$

尽管中位数对异常值的耐抗性比均值更强，但是，如果 X 为正态分布，中位数就不如均值有效。因为其他目标函数与对异常值的耐抗性加在一起，其有效稳健性就大大提高了。我们将可以最小化目标函数 $\sum_{i=1}^{n}\rho(E_i)$ 的估计值称为"M 估计量"。[46]

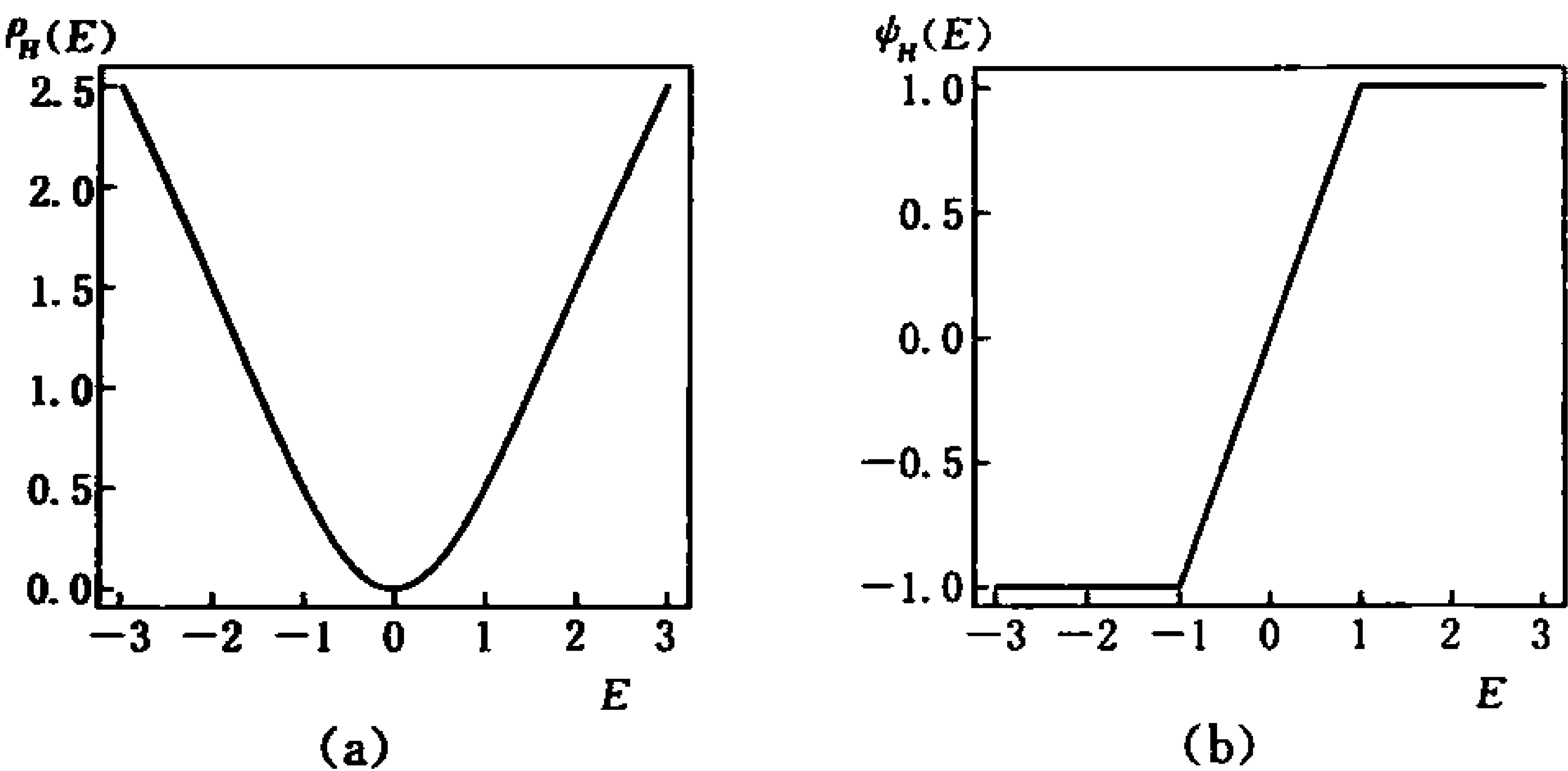
两个常见的 M 估计量是 Huber 估计量和双权或者双平方估计量。Huber 估计量是以发现 M 估计量的彼得·J. 胡

伯尔(Peter J. Huber)命名的；双权估计是由约翰·W. 杜克(John W. Tukey)——一个为统计学作出了重大贡献(其中包括稳健估计)的著名美国统计学家发明的。

Huber 估计量是最小二乘与最小绝对值之间的权衡，其数据的中心靠近最小二乘，而尾部与最小绝对值相似。

$$\rho_H(E) \equiv \begin{cases} \frac{1}{2}E^2 & (|E| \leqslant k) \\ k|E| - \frac{1}{2}k^2 & (|E| > k) \end{cases}$$

图 3.22 描述了 Huber 目标函数 ρ_H、ρ_H 的导数和影响函数 ψ_H。[47]



注：要校准这两张图，细调常数需设置为 $k=1$（请见文中有关细调常数的讨论）。

图 3.22 Huber 目标函数 ρ_H(a)和影响函数 ψ_H(b)

$$\psi_H(E) = \begin{cases} k & (E > k) \\ E & (|E| \leqslant k) \\ -k & (E < -k) \end{cases}$$

在这里，定义分布中心及尾部的 k 值称为“细调常数”。我们常常把细调常数表达成多尺度变量 X(如展宽)，即取 $k = cS$，

其中,S 是尺度的量度。样本标准差是一个不太好的尺度量度,因为它被异常值影响的程度比均值大。常见的尺度稳健量度是中位数绝对偏差(MAD):

$$\mathrm{MAD} \equiv \mathrm{median}|X_i - \hat{\mu}|$$

最初,我们用变量 X 的中位数作为 $\hat{\mu}$ 估计,然后我们定义 $S \equiv \mathrm{MAD}/0.6745$,其保证了当总体分布为正态分布时,$S$ 是标准差 σ 的估计。用 $k = 1.345S$(如 $1.345/0.6745 \approx$ 2MADs),相对于样本均值,在总体为正态分布时,加上总体为非正态分布时其对异常值所产生的耐抗性的情况下,它可以产生 95%的有效性。一个细调常数越小,其产生的耐抗性越大。

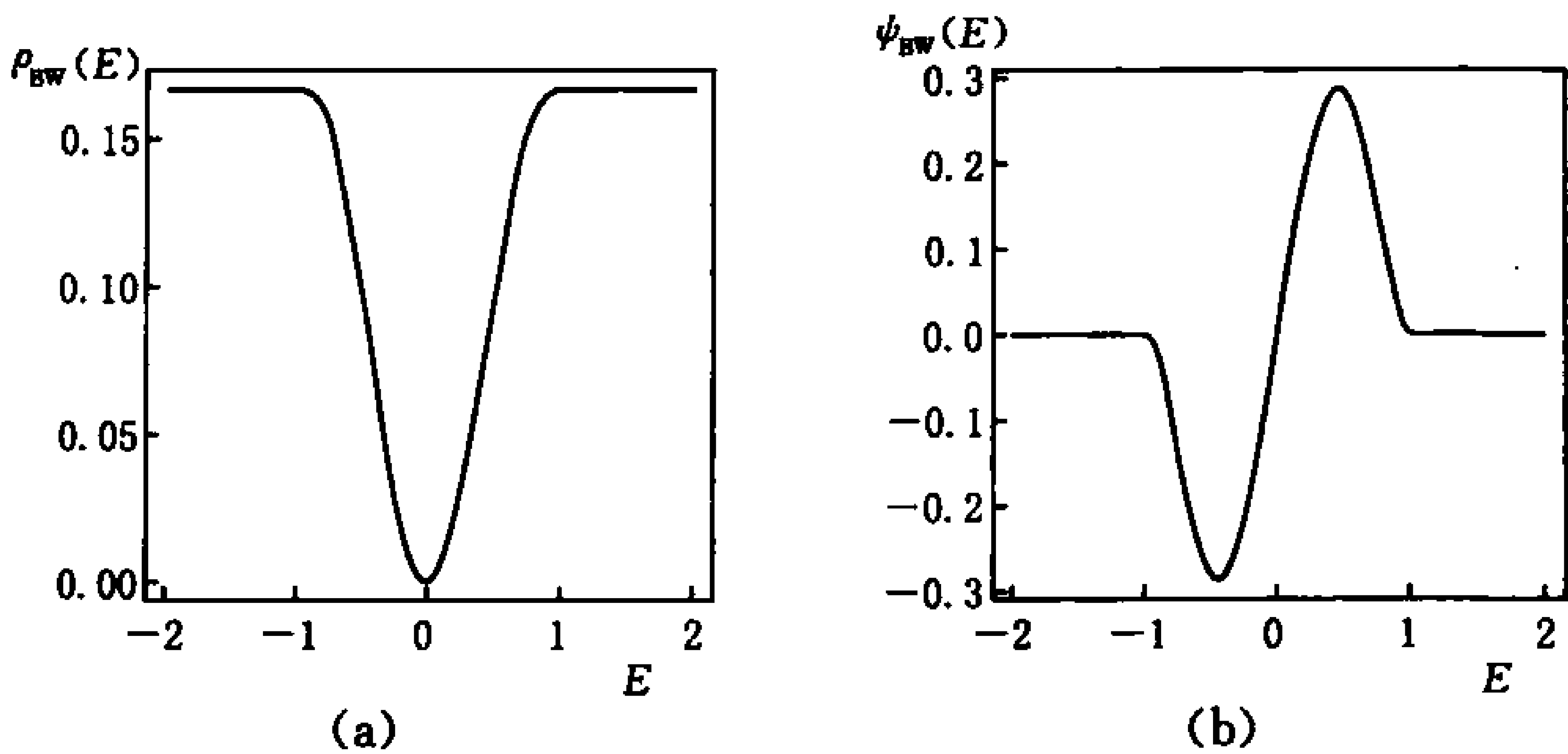
双权目标函数达到平衡或者说变平后的残差非常大[48]:

$$\rho_{\mathrm{BW}}(E) \equiv \begin{cases} \dfrac{k^2}{6}\left\{1-\left[1-\left(\dfrac{E}{k}\right)^2\right]^3\right\} & (|E| \leqslant k) \\ \dfrac{k^2}{6} & (|E| > k) \end{cases}$$

因此,双权估计量的影响函数可以再降到 0,从而完全地扣除充分异常情况的影响:

$$\psi_{\mathrm{BW}}(E) = \begin{cases} E\left[1-\left(\dfrac{E}{k}\right)^2\right]^2 & (|E| \leqslant k) \\ 0 & (|E| > k) \end{cases}$$

图 3.23 描述了有关 ρ_{BW} 和 ψ_{BW} 的函数。在样本呈正态分布时,用 $k=4.685S$(如 $4.685/0.6745 \approx 7$MADs)可以产生 95%的有效性。



注：要校准这些图形，细调常数需设定为 $k=1$。当 $|E|$ 很大时，影响函数"降"为 0。

图 3.23 双权目标函数 ρ_{BW}(a)和"影响函数"ψ_{BW}(b)

当 MAD 用于估计尺度时，Huber 和双权估计值均可实现崩溃点为 50%。

计算 M 估计值通常需要用到迭代(尽管对于均值和中位数，迭代并不是必须的，然而正如我们所见，其与 M 估计的框架相适应)。$\hat{\mu}$ 的估计方程式把目标函数的差异设置为 0，因此有：

$$\sum_{i=1}^{n} \psi(X_i - \hat{\mu}) = 0 \qquad [3.14]$$

方程 3.14 有许多解法，其中最直接、最简单的要数用迭代法对均值再加权，其过程为：

首先，定义权方程 $\omega(E) \equiv \psi(E)/E$，那么，估计方程变为：

$$\sum_{i=1}^{n} (X_i - \hat{\mu})\omega_i = 0 \qquad [3.15]$$

其中，

$$\omega_i = \omega(X_i - \hat{\mu})$$

方程 3.15 的解是加权后的均值，为：

$$\hat{\mu} = \frac{\sum \omega_i X_i}{\sum \omega_i}$$

加权函数对应的最小二乘、LAV、Huber 以及双权目标函数请参见表 3.1 和图 3.24。最小二乘权函数对每个观测都加了权，同时，双权对充分异常的数值赋予 0 值，LAV 和 Huber 不断趋近于 0 却无法达到 0。

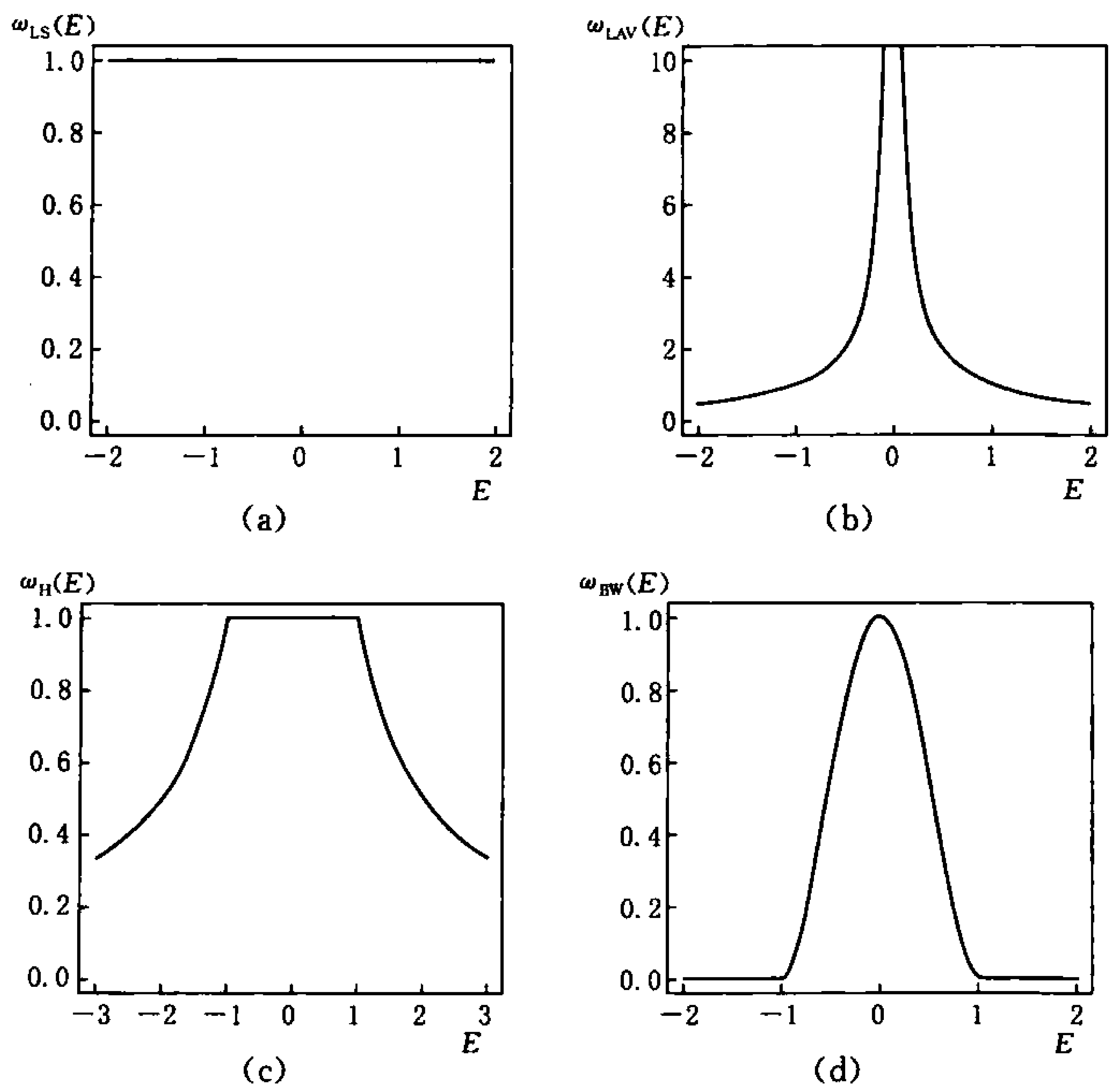
表 3.1 一些 M 估计量的权重函数

目标函数 $\rho(E)$	权重函数 $\omega(E)$
最小二乘	1
最小绝对值	$1/\vert E\vert \ (E \neq 0)$
Huber	$1 \quad (E \leqslant k)$
	$k/\vert E\vert \ (E > k)$
双权	$\left[1-\left(\frac{E}{k}\right)^2\right]^2 (\vert E\vert \leqslant k)$
	$0 \quad (E > k)$

其次，选择 $\hat{\mu}$ 的初始估计，如 X 取值的中位数。[49] 用 $\hat{\mu}^{(0)}$ 计算尺度 $S^{(0)}$ 的初始估计和初始权重 $\omega_i^{(0)} = \omega(X_i - \hat{\mu}^{(0)})$。同时，设置迭代计数初始值 $l = 0$。尺度所需的细调常数为 $k = cS$ (之前已经指定过 c)。

最后，对于每个迭代计数 l，计算 $\hat{\mu}^{(l)} = \sum \omega_i^{(l-1)} X_i / \sum \omega_i^{(l-1)}$。当从一个迭代到另一个迭代的 $\hat{\mu}^{(l)}$ 可忽略不计时，计算停止。

描述有关估计量的应用，我们首先要回顾一下之前所提到的从标准正态分布 $N(0, 1)$ 得来的含有六个观测值的样本(请见方程 3.13)，我们先在该样本中制造一个异常值



注：对于 Huber 和双权估计值，细调常数设定为 $k=1$。注意，LAV 估计量的纵坐标和 Huber 估计量的横坐标与其他图不同。

图 3.24 (a)最小二乘；(b)最小绝对值；(c)Huber；(d)双权估计量的权函数 $\omega(E)$

$X_7=10$。用 Huber 估计量 $c=1.345$ 和双权估计量 $c=4.685$，得到：

$$\bar{X}=1.298,\ \text{median}(X)=0.013,\ \hat{\mu}_H=0.201,$$

$$\hat{\mu}_{BW}=-0.161$$

很明显，样本均值已经被异常值所影响，但是其他估计量却没有。

第 6 节 | 最大似然估计

最大似然估计方法所提供的估计量是一个合理而直观的基础，它同时含有众多我们所期望的统计属性。该方法应用广泛且简单易行。再则，运用最大似然估计量，一般性理论所提供的相应的标准误和统计检验等都是有用的统计推论。然而该方法的劣势在于，它往往需要对数据结构作出较强的假定。

似然函数不仅在经典统计推论中扮演着至关重要的角色，还在贝叶斯推断中起着举足轻重的作用。

一个例子

让我们考虑一个简单的例子：假设我们要估计掷硬币得到正面的概率 π。我们投掷 10 次(例如，我们取 10 次掷硬币的结果，$n=10$)，得到的结果为：$HHTHHHTTHH$。那么，得到这个结果的概率是未知参数 π 的函数：

$$\begin{aligned}\Pr(\text{数据} \mid \text{参数}) &= \Pr(HHTHHHTTHH \mid \pi) \\ &= \pi\pi(1-\pi)\pi\pi\pi(1-\pi)(1-\pi)\pi\pi \\ &= \pi^7(1-\pi)^3\end{aligned}$$

对于10个独立的伯努利随机变量，得到该结果的概率就是每次得到正面或者反面的概率乘积(如果得到的是正面，那么，$X_i = 1$，反之 $X_i = 0$，$i = 1\cdots 10$)。

对于我们的样本，其数据是固定的，因为我们之前已经收集好了。参数 π 也有一个固定值，但是这个值是未知的，因此我们让其落在我们所想象的 0 到 1 的区间内，把观测到数据的概率看做 π 的函数。该函数为“似然函数”：

$$L(\text{数据} \mid \text{参数}) = L(\pi \mid HHTHHHTTHH) = \pi^7(1-\pi)^3$$

概率函数和似然函数的公式相同，但是概率函数是参数固定的数据函数，而似然函数是数据固定的参数函数。

下表是一些具有代表性的似然值所对应的 π 值。[50]

π	$L(\pi \mid \text{数据}) = \pi^7(1-\pi)^3$
0.0	0.0
0.1	0.0000000729
0.2	0.00000655
0.3	0.0000750
0.4	0.000354
0.5	0.000977
0.6	0.00179
0.7	0.00222
0.8	0.00168
0.9	0.000478
1.0	0.0

图 3.25 为总似然函数。尽管每个 $L(\pi \mid \text{数据})$ 的值都是一个概念上的概率，但是 $L(\pi \mid \text{数据})$ 函数既不是一个概率分

布，也不是一个密度函数，因为其总和并不为1。

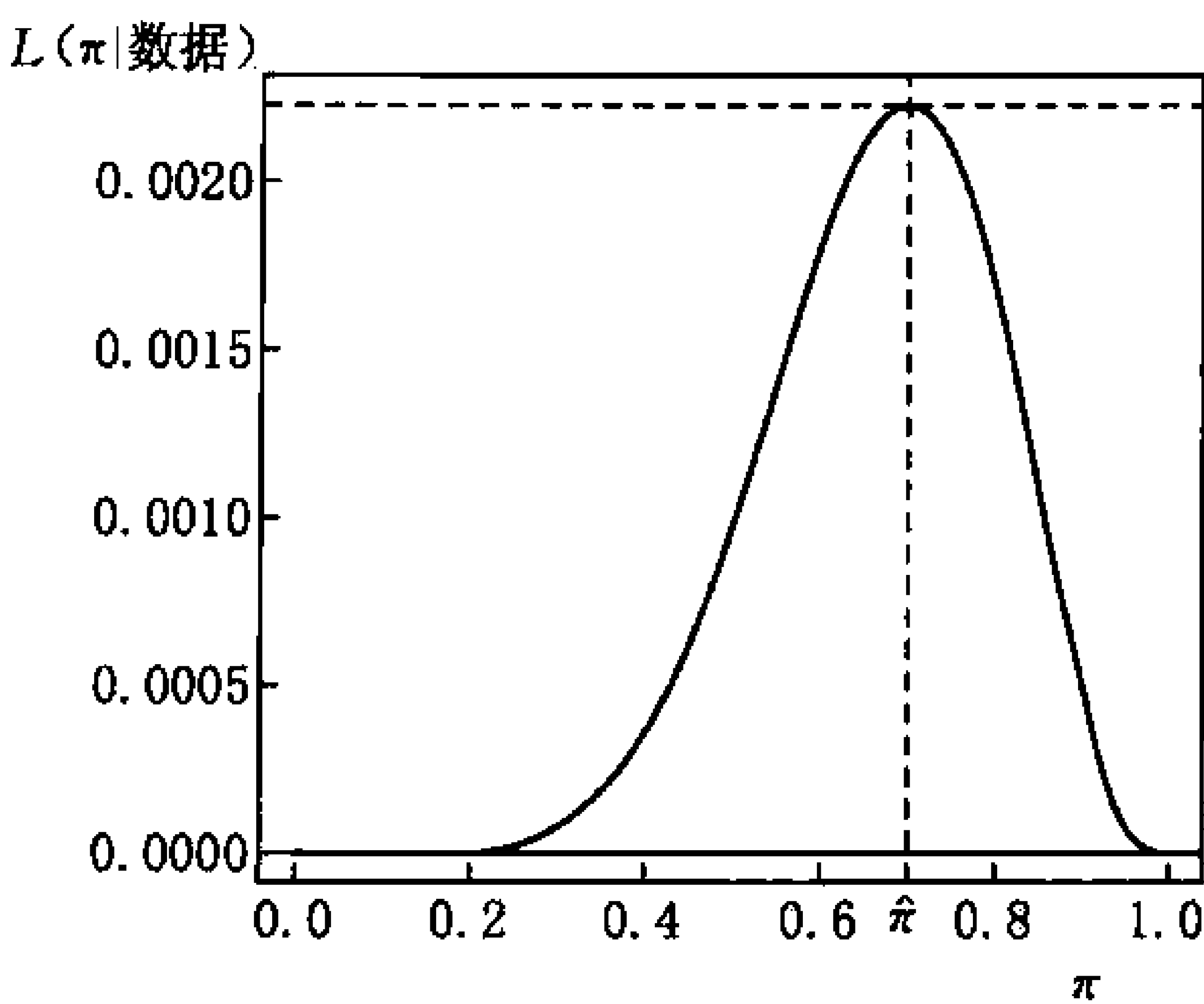


图3.25 似然函数 $L(\pi|HHTHHHTTHH)=\pi^7(1-\pi)^3$

对于此例，不论π的真实值有多大，我们已有数据样本($HHTHHHTTHH$)的概率很小。除非样本很小，否则通常任何指定的样本结果（包括我们已知的）在收集数据前得到的概率都是很小的。

尽管如此，似然函数包含了有关未知参数π的重要信息。例如，π不可能等于0或者1，因为如果它为其中任意一个值，那么，我们的观测数据（包括得到硬币的正反面）就不可能得到。反之，π值总是由数据决定，它总可以使似然函数最大化，因此，该值称为“最大似然估计”(MLE)，记做$\hat{\pi}$。在这里，$\hat{\pi}=0.7$，即得到硬币正面在样本中的比例。

将例子推广化

更普遍的情况是，我们掷硬币n次，那么得到x个正面和$n-x$个反面的概率为：

$$L(\pi \mid 数据) = \Pr(数据 \mid \pi) = \pi^x(1-\pi)^{n-x}$$

我们想得到一个π，使$L(\pi|数据)$最大。对于此例，还有一个

更简单而且等价的方法，即找到一个 π 值，使似然函数的对数最大化，这样，我们有：

$$\log_e L(\pi) = x\log_e \pi + (n-x)\log_e(1-\pi) \qquad [3.16]$$

$\log_e L(\pi)$对 π 求导得：

$$\frac{d\log_e L(\pi)}{d\pi} = \frac{x}{\pi} + (n-x)\frac{1}{1-\pi}(-1)$$
$$= \frac{x}{\pi} - \frac{n-x}{1-\pi}$$

对数似然函数对参数求导后所得的函数称为"记分"（或者"记分函数"）。将记分设置为 0 求解 π，可以得到 MLE，解方程后我们发现，MLE 即样本比例 x/n（读者可以自己证明），最大似然估计量是 $\hat{\pi} = X/n$。要避免最后阶段对估计量的替换，我们可以在对数似然函数中用 x 代替 X（如方程 3.16）。

最大似然估计量

最大似然估计量的性质如下：

（1）最大似然估计量是一致的。

（2）最大似然估计量是渐近无偏的，尽管在有限样本里它可能有偏。

（3）最大似然估计量是渐近有效的——渐近无偏估计量的渐近方差较大。

（4）最大似然估计量是正态分布的。

（5）如果一个参数含有充分统计量，那么，该参数的最大似然估计量是其充分统计量的函数。

（6）如果 $\hat{\alpha}$ 是 α 的 MLE，且 $\beta = f(\alpha)$ 是 α 的函数，那么，

$\hat{\beta} = f(\hat{\alpha})$ 是 β 的 MLE。

(7) 参数 α 的 MLE $\hat{\alpha}$ 的渐近抽样方差可以从对数似然函数的二阶导数中求得：

$$\upsilon(\hat{\alpha}) = \frac{1}{-E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]} \qquad [3.17]$$

$\upsilon(\hat{\alpha})$ 的分母称为“期望信息”或者“Fisher 信息”[51]：

$$I(\alpha) \equiv -E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]$$

我们将 MLE $\hat{\alpha}$ 代入方程 3.17，可以得到渐近抽样方差的估计 $\hat{\upsilon}(\hat{\alpha})$。[52]

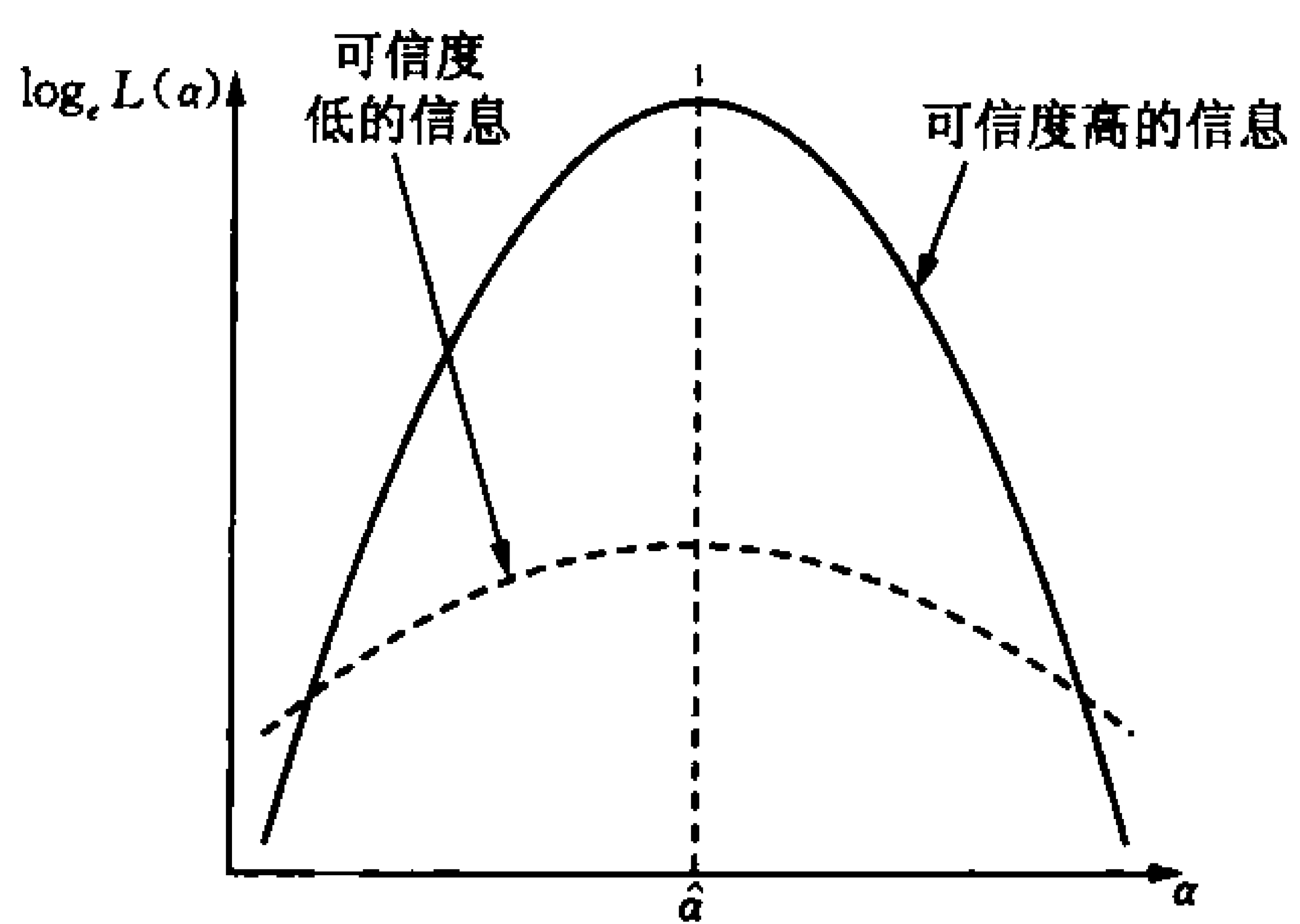
(8) $L(\hat{\alpha})$ 是似然函数在 MLE $\hat{\alpha}$ 上的值，此时，$L(\alpha)$ 对于真(但往往是未知的)参数 α 是一个似然函数。那么，其对数似然比率统计量

$$G^2 \equiv 2\log_e \frac{L(\hat{\alpha})}{L(\alpha)} = 2[\log_e L(\hat{\alpha}) - \log_e L(\alpha)]$$

遵循自由度为 1 的渐近卡方分布。因为通过定义，MLE 在我们特定的样本中最大化了似然函数，那么，在真参数值 α 下的似然函数值通常比在 MLE $\hat{\alpha}$ 下的小(除非 α 和 $\hat{\alpha}$ 碰巧相等)。

如何构建这些结果超出了本章的范围，然而这些结果的确可以给我们带来一些直观的感觉。例如，如果对数似然函数有一个尖锐峰，那么很明显，MLE 是由其临近值求导得来的。在这种情况下，其二阶导数是一个较大的负数。我们可以发现，数据里隐藏了许多有关参数值的“信息”，MLE 的抽样方差比较小，等等。相反，如果对数似然函数

在其最大值上表现得比较平坦，那么，与 MLE 差异很大的可替换估计可能和 MLE 一样好用。这样的话，数据中就很难发现有关参数值的"信息"，同时，MLE 的抽样方差也很大（见图 3.26）。



注：一个为尖锐峰，提供可信度高的参数 α 信息；另一个为平坦峰，提供的参数 α 信息可信度低。

图 3.26　两个对数似然函数

统计推论：Wald 检验、似然率检验与记分检验

前面介绍的有关最大似然估计量的属性，直接引出了用来检验假设 $H_0: \alpha = \alpha_0$ 的三个常用统计量：Wald 检验、似然率检验和记分检验。记分检验有时称为"拉格朗日乘数检验"。Wald 检验和似然率检验可以用来产生 α 的置信区间。

（1）Wald 检验：根据 MLE $\hat{\alpha}$ 的渐近正态性，我们可以计算检验统计量

$$Z_0 \equiv \frac{\hat{\alpha} - \alpha_0}{\sqrt{\hat{v}(\hat{\alpha})}}$$

它在 H_0 下是以 $N(0,1)$渐近分布的。

(2) 似然率检验：运用对数似然率后，检验统计量变为：

$$G_0^2 \equiv 2\log_e \frac{L(\hat{\alpha})}{L(\alpha_0)} = 2[\log_e L(\hat{\alpha}) - \log_e L(\alpha_0)]$$

它在 H_0 下是以 χ_1^2 渐近分布的。

(3) 记分检验：我们知道 $S(\alpha) \equiv d\log_e L(\alpha)/d\alpha$ 是对数似然函数在 α 时的斜率。在 MLE 时，记分为 0：$S(\hat{\alpha}) = 0$。那么，记分统计量的表达式为：

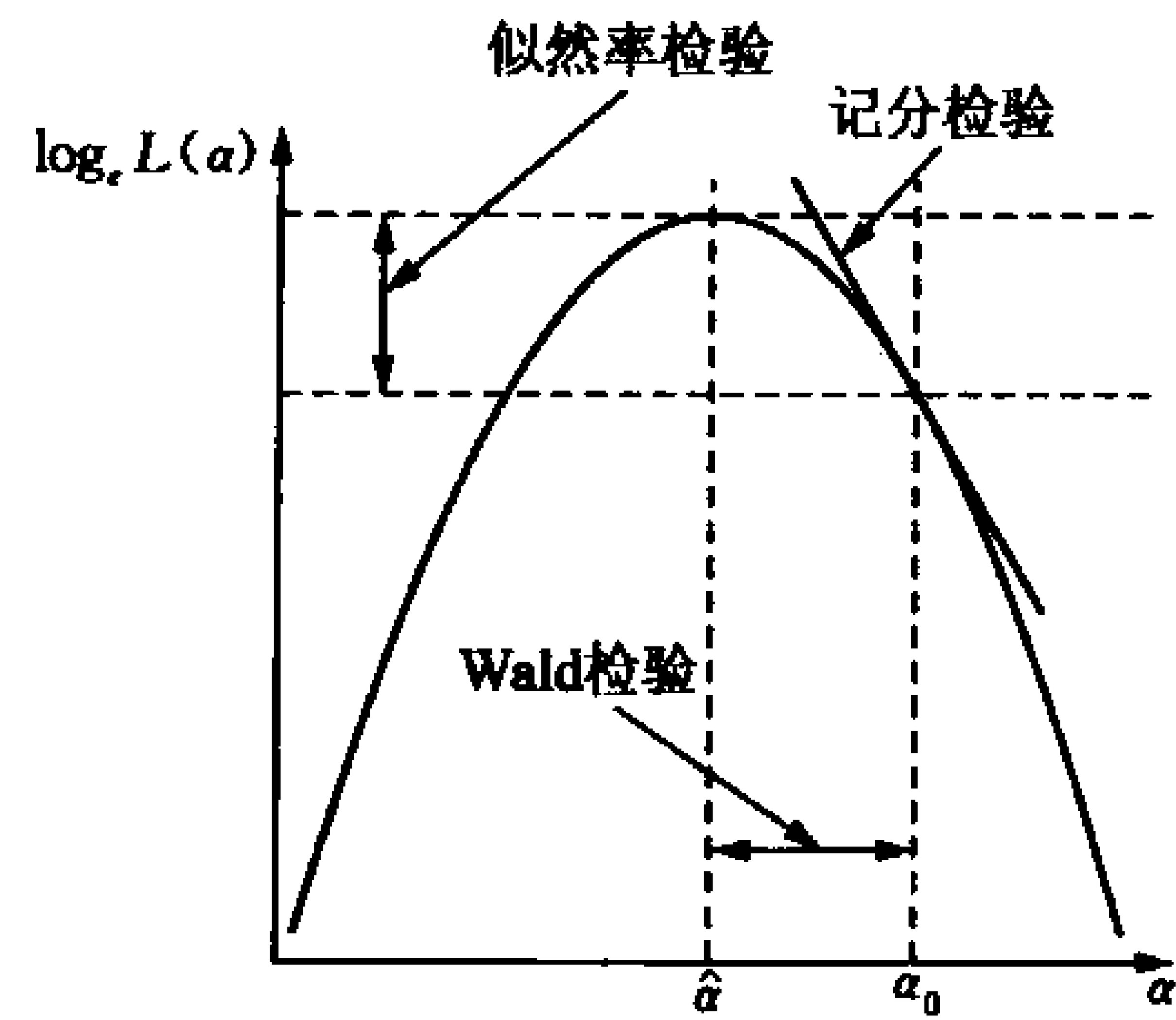
$$S_0 \equiv \frac{S(\alpha_0)}{\sqrt{I(\alpha_0)}}$$

它在 H_0 下是以 $N(0,1)$渐近分布的。

尽管这三个检验是渐近等价的，但是除非对数似然函数是二次型的，否则三个检验统计量对同一个指定样本所得到的结果会有些许不同。在特定情况下，记分检验的实际优势在于，其不需要计算 MLE$\hat{\alpha}$（因为 S_0 只依赖于空值 α_0，它已经由 H_0 指定）。在大多数小样本量的情况下，似然率检验比 Wald 检验和记分检验更可靠。

图 3.27 描述了三种检验之间的关系，并阐明了每个检验的理性直觉。Wald 检验度量了 $\hat{\alpha}$ 与 α_0 之间的距离，并用标准误校准了该距离。如果 $\hat{\alpha}$ 离 α_0 较远，那么我们可能要质疑一下 H_0。似然率检验度量了 $\log_e L(\hat{\alpha})$与 $\log_e L(\alpha_0)$之间的距离，如果 $\log_e L(\hat{\alpha})$比 $\log_e L(\alpha_0)$大得多，那么 H_0 可能出错了。记分检验的统计量度量了对数似然函数在 α_0 时的斜率，如果该斜率很陡，那么，可能离似然函数的峰值较远，

此时，我们仍然要质疑 H_0。



注：似然率检验将 $\log_e L(\hat{\alpha})$ 与 $\log_e L(\alpha_0)$ 比较；Wald 检验将 $\hat{\alpha}$ 与 α_0 比较；记分检验检验 $\alpha=\alpha_0$ 时 $\log_e L(\alpha)$ 的斜率。

图 3.27 假设检验 $H_0:\alpha=\alpha_0$

相关说明

现在我们要把这些结果运用到之前的例子中，即在 n 次掷硬币中得到正面的概率 π。之前提到，π 的 MLE 就是样本比例 $\hat{\pi}=X/n$，其中，X 记录了样本中出现正面的次数，对数似然函数的二阶求导（方程 3.16）为：

$$\frac{d^2\log_e L(\pi)}{d\pi^2}=-\frac{X}{\pi^2}-\left[-\frac{n-X}{(1-\pi)^2}(-1)\right]$$
$$=\frac{-X+2\pi X-n\pi^2}{\pi^2(1-\pi)^2}$$

注意，$E(X)=n\pi$，那么期望信息为：

$$I(\pi)=\frac{-n\pi+2n\pi^2-n\pi^2}{-\pi^2(1-\pi)^2}=\frac{n}{\pi(1-\pi)}$$

$\hat{\pi}$ 的渐近方差为 $v(\hat{\pi})=[I(\pi)]^{-1}=\pi(1-\pi)/n$，与期望信息相似。对于此例，渐近方差恰好就是 $\hat{\pi}$ 的有限样本方差，其

估计渐近抽样方差是 $\hat{v}(\hat{\pi}) = \hat{\pi}(1-\hat{\pi})/n$。

在我们的样本中，掷硬币次数 $n = 10$，得到 7 次正面的渐近抽样方差为 $\hat{v}(\hat{\pi}) = (0.7 \times 0.3)/10 = 0.0210$，根据 Wald 检验，$\pi$ 的 95%渐近置信区间为：

$$\pi = 0.7 \pm 1.96 \times \sqrt{0.02010} = 0.7 \pm 0.284$$

其中，在双尾检验中，1.96 为右侧尾部概率是 0.025 的标准正态分布值。我们还可以用 Wald 检验统计量来计算。假设 $H_0: \pi = 0.5$，

$$Z_0 = \frac{0.7 - 0.5}{\sqrt{0.02010}} = 1.38$$

其所对应的 $N(0, 1)$双尾 p 值为 0.168。

我们知道，对数似然函数为：

$$\begin{aligned}\log_e L(\pi) &= X\log_e \pi + (n - X)\log_e(1-\pi) \\ &= 7\log_e \pi + 3\log_e(1-\pi)\end{aligned}$$

代入具体数值后，得到：

$$\log_e L(\hat{\pi}) = 7\log_e L(0.7) + 3\log_e L(0.3) = -6.1086$$

$$\log_e L(\pi_0) = 7\log_e L(0.5) + 3\log_e L(0.5) = -6.9315$$

因此，H_0 的似然率检验统计量为：

$$G_0^2 = 2[-6.1086 - (-6.9315)] = 1.646$$

其所对应的 p 值(从 χ_1^2 分布得到)为 0.199。

最后，对于记分检验，

$$S(\pi) = \frac{d\log_e L(\pi)}{d\pi} = \frac{X}{\pi} - \frac{n-X}{1-\pi}$$

那么，

$$S(\pi_0)=\frac{7}{0.5}-\frac{3}{0.5}=8$$

其在 π_0 时的期望信息为：

$$I(\pi_0)=I(0.5)=\frac{10}{0.5\times 0.5}=40$$

因此记分统计量为：

$$S_0=\frac{S_{\pi_0}}{\sqrt{I(\pi_0)}}=\frac{8}{\sqrt{40}}=1.265$$

其对应的 $N(0, 1)$双尾检验的 p 值为 0.206。

这 3 个检验结果都比较一致且合理，然而却都不太准确。通过用 X 的零二项分布的精确检验(出现正面的数目)，得到：

$$p(x)=\begin{pmatrix}10\\x\end{pmatrix}0.5^{x}0.5^{10-x}=\begin{pmatrix}10\\x\end{pmatrix}0.5^{10}$$

其产生的双尾检验的 p 值为 0.3438。从这个例子中得到的经验是，在小样本量数据中应用渐近结果时一定要小心。

相关参数

最大似然方法可以推及含有多个参数的线性联立方程中，让 $p(\underset{(n\times m)}{\mathbf{X}}\mid\underset{(k\times 1)}{\boldsymbol{\alpha}})$表示 n 个可能的多元观测 $\mathbf{X}$($m\geqslant 1$)的概率或者概率密度，这些多元观测和 k 个独立参数 $\boldsymbol{\alpha}$ 有关。[53]

似然函数 $L(\boldsymbol{\alpha})\equiv L(\boldsymbol{\alpha}\mid\mathbf{X})$ 是 $\boldsymbol{\alpha}$ 的函数，此时要寻找一个 $\hat{\boldsymbol{\alpha}}$ 使得函数最大。与之前相同，我们用 $\log_e L(\boldsymbol{\alpha})$代替 $L(\boldsymbol{\alpha})$。要最大化似然函数，我们要先计算出向量偏导$\partial\log_e L(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$

并使其为0,然后来解矩阵方程求 $\hat{\boldsymbol{\alpha}}$。如果解方程结果发现有不止一个根,这时,我们就要选那个可以产生最大似然值的根。

与单个参数的例子一样,基于充分统计量的条件,最大似然估计量是一致、渐近无偏、渐近有效且为渐近正态分布的。MLE 的渐近方差一协方差矩阵为:

$$\underset{(k\times k)}{v(\hat{\boldsymbol{\alpha}})} = \left\{-E\left[\frac{\partial^2 \log_e L(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right]\right\}^{-1} \qquad [3.18]$$

方程 3.18 中括号里的矩阵称为"$I(\boldsymbol{\alpha})$"(不要和单位矩阵 **I** 混淆)。[54]另外,如果 $\beta = f(\alpha)$,那么,β 的 MLE 为 $\hat{\beta} = f(\hat{\alpha})$。注意,类比多参数方程和单一参数方程。

以下为记分检验和 Wald 检验的推广。H_0 的 Wald 统计量在 $\alpha = \alpha_0$ 时为:

$$Z_0^2 \equiv (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)'\hat{v}(\hat{\boldsymbol{\alpha}})^{-1}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$$

记分向量为 $S(\boldsymbol{\alpha}) \equiv \partial\log_e L(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$,那么,记分统计量为:

$$S_0^2 \equiv S(\boldsymbol{\alpha}_0)'I(\boldsymbol{\alpha}_0)^{-1}S(\boldsymbol{\alpha}_0)$$

似然率检验可以直接推广为:

$$G_0^2 \equiv 2\log_e\left[\frac{L(\hat{\boldsymbol{\alpha}})}{L(\boldsymbol{\alpha}_0)}\right]$$

这三个检验统计量都在 H_0 下遵循渐近分布(如 χ_k^2)。

每种检验都应该适应相对更复杂的假设。例如,我们想检验假设 H_0,其 $\boldsymbol{\alpha}$ 中 k 个元素的 p 和某个特定值相等。我们让 $L(\hat{\boldsymbol{\alpha}}_0)$代表在某些假设限制下的最大似然函数(例如,设一系列参数 p 与某假设的一系列数值相等,但其他的参数

可任意估计)；$L(\hat{\boldsymbol{\alpha}})$表示放开限制后整体的最大似然函数。那么，在 H_0 假设下，

$$G_0^2 \equiv 2\log_e\left[\frac{L(\hat{\boldsymbol{\alpha}})}{L(\boldsymbol{\alpha}_0)}\right]$$

它是自由度为 k 的渐近卡方分布。

下一个例子(Theil, 1971: 389—390)阐明了以下结果：对于一个从均值为 μ、方差为 σ^2 的正态分布得来的 n 个独立观测样本 X_i，我们想估计 μ 和 σ^2。已知似然函数为：

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(X_i-\mu)^2}{2\sigma^2}\right]$$

$$= (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right]$$

那么，其对数似然函数为：

$$\log_e L(\mu, \sigma^2) = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(X_i-\mu)^2$$

其偏导为：

$$\frac{\partial\log_e L(\mu, \sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\sum(X_i-\mu)$$

$$\frac{\log_e L(\mu, \sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(X_i-\mu)^2$$

令偏导等于 0，求得 μ、σ^2 的估计量分别为：

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X}$$

$$\hat{\sigma}^2 = \frac{\sum(X_i-\bar{X})^2}{n}$$

对数似然函数的二阶偏导矩阵为：

$$\begin{bmatrix} \frac{\partial^2 \log_e L}{\partial \mu^2} & \frac{\partial^2 \log_e L}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \log_e L}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \log_e L}{\partial (\sigma^2)^2} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum(X_i-\mu) \\ -\frac{1}{\sigma^4}\sum(X_i-\mu) & \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum(X_i-\mu)^2 \end{bmatrix}$$

取期望值会得到期望信息矩阵的负阵，注意，$E(X_i-\mu)=0$，$E(X_i-\mu)^2=\sigma^2$，

$$-I(\mu,\sigma^2)=\begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix}$$

我们知道，最大似然估计量的渐近方差—协方差矩阵即其信息矩阵的逆阵：

$$v(\hat{\mu},\hat{\sigma}^2)=[I(\mu,\sigma^2)]^{-1}=\begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

$\hat{\mu}=\overline{X}$ 的抽样方差为(σ^2/n)。σ^2 的 MLE 虽然有偏，但却是一致的(即方程 3.12 中的估计量 S_*^2)。

在许多应用中，数据包含了一个含有 n 个同分布观测的独立随机样本。数据整体的似然函数为所有观测的似然函数乘积 $L_i(\boldsymbol{\alpha})$，那么数据整体的对数似然函数则为所有观测的对数似然函数之和：

$$\log_e L(\boldsymbol{\alpha})=\sum_{i=1}^{n}\log_e L_i(\boldsymbol{\alpha})$$

因此，记分函数为逐个观测相关项之和：

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} S_i(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{\partial \log_e L_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$$

最后，样本信息是一个个体观测中所包含的 n 倍信息（记做 I_1）：

$$I(\boldsymbol{\alpha}) = nI_1(\boldsymbol{\alpha}) = nE\left[\frac{\partial^2 \log_e L_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'}\right]$$

结果之所以如此，是因为似然函数的二阶导数对 n 个观测都是相同的。

Delta 算法

如前所述，假如 $\beta = f(\alpha)$，且 $\hat{\alpha}$ 为 α 的最大似然估计量，那么，$\hat{\beta} = f(\hat{\alpha})$ 为 β 的最大似然估计量。这意味着 $\hat{\beta}$ 是渐近正态分布的，且其渐近期望值为 β，即使函数 $f(\cdot)$ 是非线性的。

利用 $f(\hat{\alpha})$ 泰勒展式估计在 α 处一阶展开，Delta 算法产生了一个 $\hat{\beta}$ 渐近方差的估计：

$$\hat{\beta} = f(\hat{\alpha}) \approx f(\alpha) + f'(\alpha)\left(\hat{\alpha} - \alpha\right) \qquad [3.19]$$

其中，$f'(\alpha) = df(\alpha)/d\alpha$ 为 $f(\alpha)$ 对 α 的求导。

方程 3.19 右边的 $f(\alpha)$ 是一个常数（因为参数 α 是定值），第二项是关于 $\hat{\alpha}$ 的线性函数，由于 α 为定值，因此 $f'(\alpha)$ 为常数，所以，

$$v\left(\hat{\beta}\right) \approx \left[f'\left(\hat{\alpha}\right)\right]^2 v\left(\hat{\alpha}\right)$$

其中，$v\left(\hat{\alpha}\right)$为 $\hat{\alpha}$ 的渐近方差。在实际应用中，我们用最大似然估计量 $\hat{\alpha}$ 代替 α，进而获得 $\hat{\beta}$ 的渐近方差估计值：

$$\hat{v}\left(\hat{\beta}\right)=\left[f'\left(\hat{\alpha}\right)\right]^2 v\left(\hat{\alpha}\right)$$

为了解释 Delta 算法的应用，让我们先回顾一些概念，样本配比 $\hat{\pi}$ 为总体配比 π 的最大似然估计量，其渐近（实际上是有限样本）方差为 $v\left(\hat{\pi}\right)=\pi(1-\pi)/n$，其中，$n$ 为样本大小。对数优比或者 logit 的定义为：

$$\Lambda=f(\pi)\equiv\log_e\frac{\pi}{1-\pi}$$

Λ 的最大似然估计量为 $\hat{\Lambda}=\log_e[\hat{\pi}/(1-\hat{\pi})]$，logit 样本的样本渐近方差为：

$$\begin{aligned}v(\hat{\Lambda})&\approx[f'(\pi)]^2 v(\hat{\pi})\\&=\left[\frac{1}{\pi(1-\pi)}\right]^2\frac{\pi(1-\pi)}{\pi}\\&=\frac{1}{n\pi(1-\pi)}\end{aligned}$$

最后，logit 样本的样本方差渐近估计值为 $\hat{v}\left(\hat{\Lambda}\right)=1/\left[n\hat{\pi}\left(1-\hat{\pi}\right)\right]$。

Delta 算法可以直接扩展到具有多个参数的函数中。假设 $\beta=f(\alpha_1,\alpha_2,\cdots,\alpha_k)=f(\alpha)$，且 $\hat{\alpha}$ 为 α 的最大似然估计量，其渐近方差为 $v\left(\hat{\alpha}\right)$，那么，$\hat{\beta}=f\left(\hat{\alpha}\right)$ 的渐近方差为：

$$v\left(\hat{\beta}\right)\approx[\mathbf{g}(\boldsymbol{\alpha})]'v\left(\hat{\boldsymbol{\alpha}}\right)\mathbf{g}(\boldsymbol{\alpha})=\sum_{i=1}^{k}\sum_{j=1}^{k}v_{ij}\times\frac{\partial\hat{\beta}}{\partial\alpha_i}\times\frac{\partial\hat{\beta}}{\partial\alpha_j}$$

其中，$\mathbf{g}(\boldsymbol{\alpha}) = \partial\hat{\beta}/\partial\boldsymbol{\alpha}$，$\upsilon_{ij}$ 为$\upsilon(\hat{\alpha})$的第 i 行第 j 列元素。$\hat{\beta}$ 的估计渐近方差为：

$$\hat{\upsilon}\left(\hat{\beta}\right) = \left[\mathbf{g}\left(\hat{\boldsymbol{\alpha}}\right)\right]'\upsilon\left(\hat{\boldsymbol{\alpha}}\right)\mathbf{g}\left(\hat{\boldsymbol{\alpha}}\right)$$

Delta 算法不仅适用于最大似然函数的估计量，而且适用于其他渐近正态分布的估计量。

第 7 节 | 贝叶斯推断

本章节引入另外一种统计推断，即贝叶斯推断。这里主要解释贝叶斯推断的核心思想，细节的部分将会被省去。

贝叶斯定理

首先，我们来回顾条件概率的定义。已知事件 B 会发生，A 发生的概率为：

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \qquad [3.20]$$

同样，B 关于 A 的条件概率为：

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} \qquad [3.21]$$

将方程 3.21 进行变换，得到 A 和 B 的联合概率：

$$\Pr(A \cap B) = \Pr(B \mid A)\Pr(A)$$

并将其代入方程 3.20，可得到贝叶斯定理：

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \qquad [3.22]$$

贝叶斯定理是以 18 世纪英国数学家托马斯 · 贝叶斯(Thom-

as Bayes)的名字命名的。

贝叶斯统计推断是基于方程 3.22 的推断。让 A 代表某未知命题,我们想弄清楚其正确与否(例如这样一个命题:一个参数等于某特定值)。让 B 代表与真命题相关的观测数据。无条件概率 $\Pr(A)$为 A 的先验概率,它是在获得数据之前,我们确信 A 为真的概率;$\Pr(B|A)$是假设 A 为真时获得观测数据的概率,即给定 A 的似然度。数据 B 的无条件概率为:

$$\Pr(B) = \Pr(B \mid A)\Pr(A) + \Pr(B \mid \bar{A})\Pr\left(\bar{A}\right)$$

那么,方程 3.22 中的 $\Pr(A|B)$就是 A 的后验概率,表示获得数据 B 后所修正的 A 为真的概率。

贝叶斯推断是基于证据、检验先验的理性过程。主观论者和客观论者关于概率的理解是相反的。贝叶斯理论由初等概率理论发展而来,随后便引出了贝叶斯统计推断的一般过程。

初步案例

考虑如下的简单情况:假设你有两个"有偏差的"硬币,在抛掷过程中,其中一个得到正面的概率为 $\Pr(H) = 0.3$,另一个为 $\Pr(H) = 0.8$。每个硬币都分别被装在盒子里,且在盒子上标明了它的偏差。但是你不小心把盒子都弄丢了,只好把硬币都放在抽屉里。一年后,你忘记了哪个硬币是哪个。为了区分它们,你随便挑了一个,然后抛了 10 下,得到结果为 $HHTHHHTTHH$——七次正面,三次反面。

让事件 A 代表选取了硬币 $\Pr(H) = 0.3$,那么 $\bar{A}$ 为事件选取 $\Pr(H) = 0.8$。在这种情况下,我们有理由选择先验

概率 $\Pr(A)=\Pr(\bar{A})=0.5$，那么数据的似然度为：

$$\Pr(B \mid A)=0.3^{7}(1-0.3)^{3}=0.0000750$$

$$\Pr(B \mid \bar{A})=0.8^{7}(1-0.8)^{3}=0.0016777$$

请注意，常见观测数据的似然度在两种情况下都很小，但是 $\bar{A}$ 的情况更有可能。

利用贝叶斯定理(方程 3.22)，我们可得到后验概率：

$$\Pr(A \mid B)=\frac{0.0000750\times 0.5}{0.0000750\times 0.5+0.0016777\times 0.5}=0.0428$$

$$\Pr(\bar{A} \mid B)=\frac{0.0016777\times 0.5}{0.0000750\times 0.5+0.0016777\times 0.5}=0.9572$$

此结果说明，所选的硬币为 $\Pr(H)=0.8$ 的概率比 $\Pr(H)=0.3$ 的概率更大。

贝叶斯定理扩展

贝叶斯定理可以轻易地扩展到多于两个假设 A 和 $\bar{A}$ 的情况。比如有多个假设 $H_1, H_2, \cdots, H_k$，其先验概率分布为 $\Pr(H_i), i=1, \cdots, k$，且所有先验概率的和为 1[55]；让 D 代表观测的数据，并有似然度 $\Pr(D|H_i), i=1, \cdots, k$，那么，假设 H_i 的后验概率为：

$$\Pr(H_i \mid D)=\frac{\Pr(D \mid H_i)\Pr(H_i)}{\sum_{j=1}^{k}\Pr(D \mid H_j)\Pr(H_j)} \qquad [3.23]$$

方程 3.23 的分母确保了在所有假设下，后验概率的和为 1。有时候，为方便起见，我们可以省略这个标准化，将其简单表示为：

$$\Pr(H_i \mid D) \propto \Pr(D \mid H_i)\Pr(H_i)$$

即一个假设的后验概率与该假设下的似然度和其先验概率的乘积成正比。如果有必要，我们可以除以 $\sum \Pr(D \mid H_i)\Pr(H_i)$ 来复原后验概率。

贝叶斯定理对于随机变量同样适用。让 α 代表我们感兴趣的参数，它的先验概率分布或者密度为 $p(\alpha)$；让 $L(\alpha) \equiv p(D \mid \alpha)$ 表示参数 α 的似然函数，那么有：

$$p(\alpha \mid D) = \frac{L(\alpha)p(\alpha)}{\sum_{\text{all}\alpha'} L(\alpha')p(\alpha')}$$

其中，α 是离散的，或者

$$p(\alpha \mid D) = \frac{L(\alpha)p(\alpha)}{\int L(\alpha')p(\alpha')d\alpha'}$$

因为在更普遍的情况下，α 是连续的。在两种条件下都有：

$$p(\alpha \mid D) \propto L(\alpha)p(\alpha)$$

即后验分布或者密度与似然函数和先验概率（或者密度）的乘积成正比。跟前面一样，如果有需要，我们可以除以 $\sum_{\text{all}\alpha'} L(\alpha')p(\alpha')$ 或者 $\int L(\alpha')p(\alpha')d\alpha'$ 来复原后验概率或者密度。

有两点需要提及：

首先，进行贝叶斯推断之前，我们要求参数 α 的先验分布 $p(\alpha)$是合理的。

另外，与经典统计量相反，我们把 α 当做一个随机变量而不是未知常数，所以我们保留希腊字母。然而，由于与数据不同，参数永远不能确定——即使已经获得了数据。

共轭先验

当先验分布已经选定，且似然函数和先验概率的乘积所得到的后验分布与该先验分布属于同一个系列，此时贝叶斯推断的数学会变得很简单。我们把这种情况的先验分布叫做“共轭先验”。

贝叶斯推断曾经只在共轭先验的情况下才有实用价值。然而，随着计算机软件和硬件的发展，通过随机取样，数学上难以解决的后验分布变为可能。比如，马尔科夫链蒙特卡罗抽样(Markov Chain Monte Carlo，MCMC)使得贝叶斯能广泛应用于统计学。但是不论怎样，先验分布的选择是非常重要的。

贝叶斯推断的例子

让我们继续之前的例子——掷硬币，我们想通过估计得到硬币正面的概率 π，但是在少量离散值中又无法限制 π。原则上，π 可以为 0 到 1 之间的任意数值。要估计 π，需要收集 10 次独立投掷的数据。从之前的伯努利似然函数中，我们知道：

$$L(\pi) = \pi^h (1-\pi)^{10-h} \qquad [3.24]$$

其中，h 为观测到的出现硬币正面的次数。通过实验，我们得到数据 $HHTHHHTTHH$，因此，$h = 7$。

方程 3.24 伯努利似然函数的共轭先验即贝塔分布，

$$p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,\ b)} \quad (0 \leqslant \pi \leqslant 1 \text{ 且 } a,\ b \geqslant 0)$$

当贝塔先验与似然函数相乘后，我们得到了一个后验密度

形式：

$$p(\pi \mid D) \propto \pi^{h+a-1}(1-\pi)^{10-h+b-1} = \pi^{6+a}(1-\pi)^{2+b}$$

即贝塔分布的形状参数为 $h+a=7+a$，$10-h+b=3+b$。在效果上，先验概率在似然函数里将 a 和 b 分别加到了出现正面和反面的次数中。

那么，我们应该如何选择 a 和 b 呢？方法之一可以反映你对似然值 π 的主观估计。例如，如果一个硬币本身没有问题，那么，π 的值会很接近 0.5。假设取 $a=b=16$，那么，π 就会被限制在 0.3 到 0.7 的区间内（见图 3.15）。如果对该限制不太满意，对于 a 和 b，我们可以取小一点的值。当 $a=b=1$ 时，π 的所有值都可能相等，这就是所谓的“扁平先验分布”，它完全忽略了 π 值。[56]

图 3.28 描述了 π 在两种先验下的后验分布。在扁平先验下，后验和似然函数成正比，因此，如果我们取后验的众数作为 π 的估计，就会得到 MLE$\hat{\pi}=0.7$。[57]相反，对于 $a=b=16$ 这个信息先验，其众数在 $\pi\approx 0.55$，它和 $\pi=0.5$ 的先验分布的众数非常接近。

令人不安的是，该结论要取决于关键的先验分布，但这个结果却是在少量的样本下得到的。我们知道，在这种情况下用贝塔先验如同在数据中又增添了 $a+b$ 个观测。随着样本量的增加，似然函数开始占领后验分布，先验分布逐渐被掩盖。[58]对于此例，如果掷 n 次硬币，其后验分布的形式变为：

$$p(\pi \mid D) \propto \pi^{h+a-1}(1-\pi)^{n-h+b-1}$$

得到的 h 个正面和 $n-h$ 个反面的数目都会随着掷币次数的增加而增加。直觉告诉我们，从先验来说，样本量小时所要

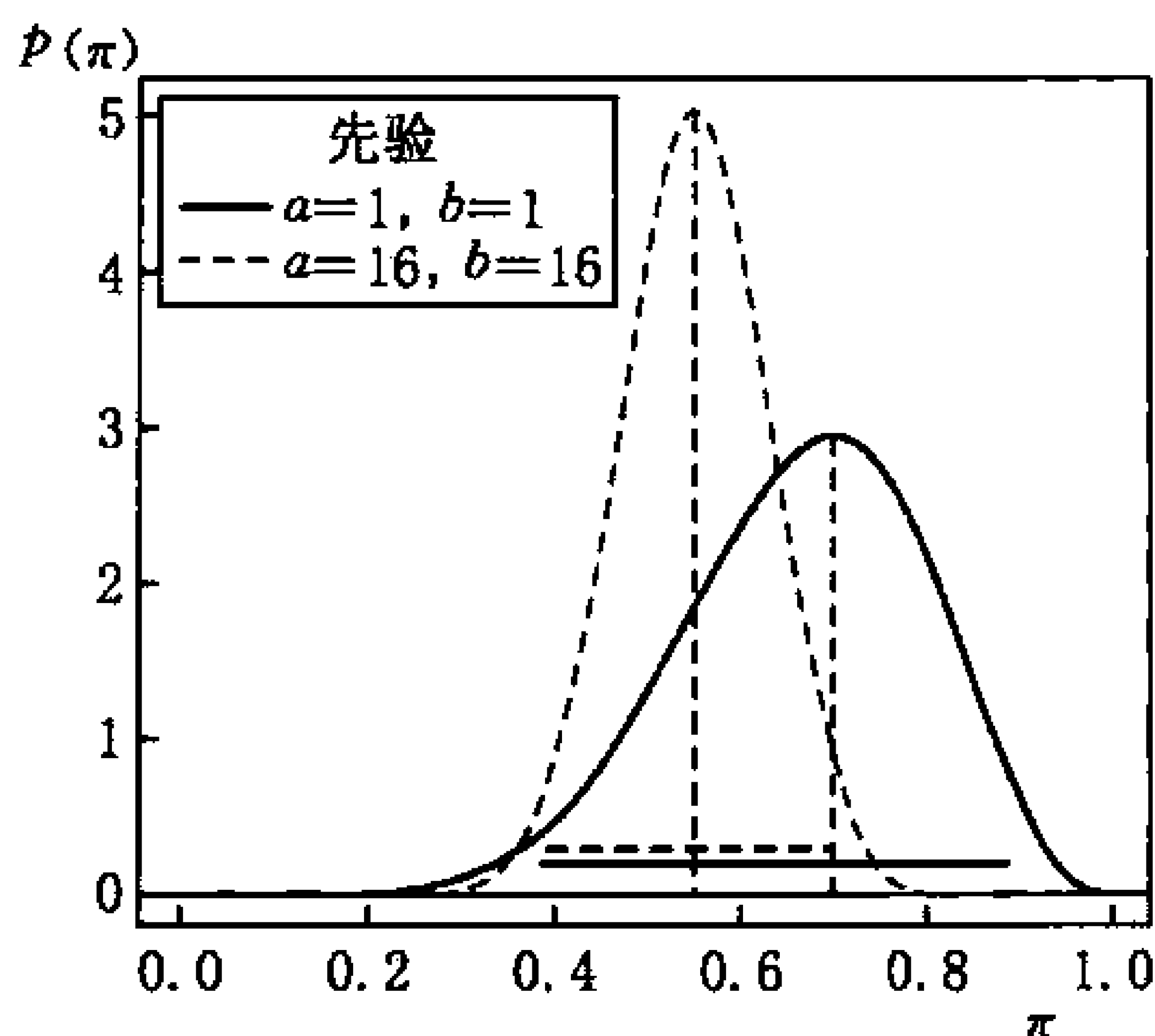
考虑的加权要比样本量大的时候大得多。

贝叶斯区间估计

在经典统计推理中，我们期望得到的不仅是参数的一个点估计，还要检验估计中的不确定性。参数的后验分布直接表示了统计不确定性。通过后验分布，我们可以进行多种贝叶斯区间估计，这些区间估计均可以用来和经典置信区间进行对比。

一个简单的选择是中央后验区间：总共含 $100a$ 百分比的中央后验区间为从 $(1-a)/2$ 到 $(1+a)/2$ 的分位数。与以解释复杂而著名的经典置信区间不同，贝叶斯后验区间的解释很简单：概率是 0.95 表示其参数落在 95%的置信区间内。该差别反映了贝叶斯参数解释是把参数当做一个随机变量，对数据进行观测后，其通过后验分布表达了对参数值的主观不确定。

两个后验分布的 95%中央后验区间可从图 3.28 中看出。



注：$a=1$，$b=1$ 是扁平的贝塔先验；$a=16$，$b=16$ 是信息贝塔先验。其中，在 10 次掷币中包含七次正面。两个靠近图底部的水平线分别展示了相应先验的 95%中央后验区间。

图 3.28　在两个先验分布下得到正面概率 π 的后验分布

贝叶斯参数推理

贝叶斯推理可以直接延伸到对一系列参数 $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_k]'$ 的同时估计中。在这种情况下，有必要指定参数的联合先验分布 $p(\boldsymbol{\alpha})$ 和联合似然估计 $L(\boldsymbol{\alpha})$。那么，对于只有单一参数的情况，联合后验分布与先验分布和似然估计的乘积成正比：

$$p(\boldsymbol{\alpha} \mid D) \propto p(\boldsymbol{\alpha})L(\boldsymbol{\alpha})$$

该推理主要关注每个参数的边缘后验分布 $p(\alpha_i \mid D)$。

第8节 | 推荐阅读

大多数介绍性的数理统计和计量经济学教材都会涵盖本章所提及的各个主题，且其描述更为正式和详尽。如考克斯(Cox)和欣克利(Hinkley)(1974)的著作，相对于本书，其涉及的知识较难。还有策尔纳(Zellner)(1983)的著作，其结构紧凑，与考克斯(Cox)和欣克利(Hinkley)(1974)的书相比，较为简单。旺纳科特兄弟的(Wonnnacott & Wonnnacott, 1990)著作用相对简单的数学知识对本章涉及的话题进行了深刻诠释。如果你觉得本章相关章节过于精炼且缺少细节，那么我认为这本书无疑非常合适。同时，关于渐近分布理论还有泰尔(Theil)(1971)的著作。关于 Wald 检验、似然率和记分检验还可参看英格尔(Engle)(1984)的著作。最后，对于贝叶斯推理的相关内容，兰开斯特(Lancaster)(2004)的书确实是一本经典之作。

第4章

实际应用：线性最小二乘法回归

这本书的重点在于介绍社会统计学的数学方法，而不在于统计方法本身。不过，我还是觉得有必要介绍如何将数学应用到统计学方法中。所以，本章的目的是阐述线性最小二乘法回归这一统计方法的发展过程——一个读者所熟悉的话题以及由此推衍的相关特性。

首先，本章将描述最小二乘法的数学性质，但是这只是统计学的一部分。虽然数学在应用统计学中扮演了重要的角色，但是应用统计学不全是数学，其范围更广，例如，有关方法论的话题。此外，线性最小二乘回归在几个方面代表了应用统计学的核心方法，且在统计学中经常用到，它容易扩展到一般线性模型、广义线性模型和其他模型，并为其他统计模型提供了运算基础。最后，对于线性最小二乘法回归在数据分析中所扮演的角色，需要更深入的探讨。因此，我认为这是一本有关应用回归分析比较合适的教材（Fox, 2008）。

本章将把前面几章所学的内容应用到统计方法中：第 1 章的矩阵和线性代数，包括矩阵秩和线性联立方程；第 2 章中最优化问题所用到的矩阵微积分；第 3 章的概率论、统计分布、估计量性质和最大似然法估计。

第1节｜最小二乘法拟合

一个线性回归方程可以写成:

$$Y_i = A + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik} + E_i \qquad [4.1]$$

其中,Y_i 是 n 个观测中的第 i 个定量响应变量(或者“因变量”);x_{i1}, x_{i2}, …, x_{ik} 为第 i 个观测的 k 个定量解释变量(或者“自变量”);A, B_1, B_2, …, B_k 为回归系数,A 为回归所得的截距或者常数;系数 $B_i(i = 1, 2, \cdots, k)$ 为分项斜率系数;E_i 为回归残差,表示 Y_i 偏离线性回归面的程度。

$$\hat{Y}_i = A + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik}$$

其中,$\hat{Y}_i$为第 i 观测的拟合值。

注意,上例中我们用了大写字母 Y_i 和 E_i,这表明,如果我们所选择的(含有 n 个观测的)样本不同,因变量的值就会改变,残差也会改变。因此,Y_i 和 E_i 是随机变量。同样,因为回归系数随着样本的改变而改变,所以它们也用大写字母表示。相反,我用小写字母表示解释变量,表明在重复抽样中,它们的值是固定的,这是典型的实验设计,因为所有自变量 x 都是由研究者直接控制的,在重复实验中不会改变。把所有自变量当做定值会使数学变得简单,同样也会使之变得不太重要(但不是绝对不重要)。在下文中,我会简要地对一

系列随机变量 X 进行介绍。

我们将最小二乘法回归系数(即 A 和 B 的那些可以使残差平方和最小的取值)看做一个回归系数的函数，则有：

$$S(A, B_1, \cdots, B_k) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$
$$= \sum_{i=1}^{n} (Y_i - A - B_1 x_{i1} - \cdots - B_k x_{ik})^2$$

虽然我们可以继续用纯量形式，但是矩阵形式更有优势。我们可以把方程 4.1 改写成：

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}} \underset{(k+1\times 1)}{\mathbf{b}} + \underset{(n\times 1)}{\mathbf{e}}$$

其中，$y = [Y_1, Y_2, \cdots, Y_n]'$ 是一组观测的因变量向量，

$$\mathbf{X} \equiv \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

为模型(设计)矩阵，它包含了解释变量及首列为 1 的回归常数(常数回归因子)；$\mathbf{b} \equiv [A, B_1, \cdots, B_k]'$ 包含了回归系数；$\mathbf{e} \equiv [E_1, E_2, \cdots, E_n]'$ 为一个残差向量。那么，残差平方和为：

$$\begin{aligned} S(\mathbf{b}) &= \mathbf{e}'\mathbf{e} \\ &= (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Xb} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \end{aligned} \qquad [4.2]$$

由于 $\mathbf{y}'\mathbf{Xb}$ 为 (1×1)，因此，它和其转置 $\mathbf{b}'\mathbf{X}'\mathbf{y}$ 相等。

为最小化残差平方和 $S(\mathbf{b})$，我们可以对回归系数 $\mathbf{b}$ 求

导，将方程4.2代入可得：

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}$$

使之为0并移项，可得线性最小二乘法回归的正态方程：

$$\underset{(k+1\times k+1)}{\mathbf{X}'\mathbf{X}}\ \underset{(k+1\times 1)}{\mathbf{b}} = \underset{(k+1\times 1)}{\mathbf{X}'\mathbf{y}}$$

这是一个拥有$k+1$个线性等式和$k+1$个未知回归系数$\mathbf{b}$的系统方程。该系统方程的系数矩阵为：

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{ik} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ik} \\ \sum x_{i2} & \sum x_{i2}x_{i1} & \sum x_{i2}^2 & & \sum x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ik} & \sum x_{ik}x_{i1} & \sum x_{ik}x_{i2} & \cdots & \sum x_{ik}^2 \end{bmatrix}$$

它包含了平方和及模型矩阵的列的交叉乘积。方程右边的向量，$\mathbf{X}'\mathbf{y} = [\sum Y_i, \sum x_{i1}Y_i, \sum x_{i2}Y_i, \cdots, \sum x_{ik}Y_i]'$，包含了模型矩阵每一列交叉乘积的和以及因变量向量。平方和及乘积$\mathbf{X}'\mathbf{X}$和$\mathbf{X}'\mathbf{y}$可以由数据直接计算得到。

$\mathbf{X}'\mathbf{X}$是满秩的，即非奇异的，假如模型矩阵$\mathbf{X}$是列满秩$k+1$，则没有一个自变量是其他自变量的完美线性函数。在这些条件下，正态方程有唯一解：

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad [4.3]$$

这是$S(\mathbf{b})$的一个最小值，因为$\mathbf{X}'\mathbf{X}$是非奇异且正定的。

第 2 节 | 一个线性回归的统计模型

一个常用的线性回归统计模型为：

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

其中，Y_i 为 n 个样本观测中的第 i 个响应值；x_i，x_{i2}，…，x_{ik} 为 k 个解释变量；α，β_1，β_2，…，β_k 为总体回归系数，它是从样本数据中估算得来的；ε_i 为第 i 个观测的误差变量。即使误差不是随机变量，我们还是用希腊字母表示，因为它不能直接观测。我们假设误差是正态分布的，且其均值为 0，方差为常数 σ^2，$\varepsilon_i \sim N(0, \sigma^2)$，不同观测的误差是相互独立的。

最后，将线性方程写成矩阵形式：

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}}\underset{(k+1\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \qquad [4.4]$$

其中，$\mathbf{y}$ 为因变量，X 为模型矩阵，$\boldsymbol{\beta} = [\alpha, \beta_1, \beta_2, \cdots, \beta_k]'$ 为总体回归系数向量，$\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]'$ 为误差向量。误差向量是具有纯量协方差矩阵的多元正态分布向量，$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 I_n)$。请注意，由于它们是独立的，所以不同的误差是不相关的。[59]

因变量 $\mathbf{y}$ 的分布遵循 $\boldsymbol{\varepsilon}$ 的分布：

$$\boldsymbol{\mu} = E(\mathbf{y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

$$= \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$$

$$V(\mathbf{y}) = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})']$$

$$= E[(\mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{X}\boldsymbol{\beta})']$$

$$= E[\boldsymbol{\varepsilon\varepsilon}'] = \sigma^2 \mathbf{I}_n$$

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

因此,假设 $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ 隐含了 $E(\mathbf{y})$是 $\mathbf{X}$ 的线性函数。

第3节 | 作为估计量的最小二乘法系数

方程4.3的最小二乘法回归系数 **b** 可能可以用来估计方程4.4的线性回归模型的系数。由于 **b** 是由因变量 **y** 经过线性变换得来的，因此，最小二乘估计量的性质可以简单表现为：

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y}$$

其中，变换矩阵 $\mathbf{M} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$。因为模型矩阵 **X** 对于重复抽样是固定的，所以 **M** 亦如此。那么，

$$E(\mathbf{b}) = \mathbf{M}E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

因此可证明，**b** 为 $\boldsymbol{\beta}$ 的无偏估计量。请注意，该结论取决于假设 $E(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}$（即线性假设）。

b 的协方差矩阵可从常数误差方差和误差不相关，即 $V(\mathbf{y})=\sigma^2\mathbf{I}_n$ 的假设得来：

$$\begin{aligned} V(\mathbf{b}) &= \mathbf{M}V(\mathbf{y})\mathbf{M}' \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2\mathbf{I}_n[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

最后,根据误差正态分布假设,我们有:

$$\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}] \qquad [4.5]$$

最小二乘法估计量 b 不仅仅是 $\boldsymbol{\beta}$ 的一个无偏估计量,而且在线性、常误差方差和独立性假设下,是数据线性函数的一个最小方差无偏估计量。该结果称为"高斯-马尔科夫定理"(Gauss-Markov Theorem),常用来支持最小二乘法估计,但是不太支持最小二乘法估计量。当误差分布为非正态时,数据的其他非线性无偏估计量(所谓的稳健性回归估计量)比最小二乘估计量更为有效。但是,当误差是正态分布的时候,最小二乘法估计量将是所有无偏估计量中最有效、最可信的。[60]

第4节 | 回归模型的统计推断

有关总体回归系数 $\boldsymbol{\beta}$ 的统计推断，除了点估计外都很复杂，因为我们基本上不知道误差方差 σ^2，所以不能直接将方程4.5用于 $\mathbf{b}$ 的最小二乘估计量的分布。我们必须先估计 σ^2。

σ^2 的一个无偏估计量为：

$$S^2 = \frac{\sum_{i=1}^{n} E_i^2}{n-k-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$$

其中，$n-k-1$ 为误差的自由度(估计 $\boldsymbol{\beta}$ 的 $k+1$ 个元素时，"损失"了 $k+1$ 个自由度)。那么，估计的最小二乘协方差矩阵为：

$$\hat{V}(\mathbf{b}) = S^2(\mathbf{X}'\mathbf{X})^{-1}$$

$\mathbf{b}$ 的对角元平方根为回归系数的标准差：$SE(A)$，$SE(B_1)$，…，$SE(B_k)$。

个体回归系数的推断是建立在 t 分布上的。例如，检验零假设 $H_0: \beta_j = \beta_j^{(0)}$，即斜率系数等于一个特定值 $\beta_j^{(0)}$(一般为0)，我们可以计算检验统计量：

$$t_0 = \frac{B_j - \beta_j^{(0)}}{SE(B_j)}$$

它在零假设下是以 t_{n-k-1} 分布的。同样,建立一个 β_j 的 95%的置信区间,我们取

$$\beta_j = B_j \pm t_{n-k-1,\,0.025} SE(B_j)$$

其中,$t_{n-k-1,\,0.025}$ 是自由度为 $n-k-1$、右侧尾部概率为 0.025 的临界值。

在一般情况下,我们可以检验线性假设:

$$H_0: \underset{(q\times k+1)}{\mathbf{L}}\ \underset{(k+1\times 1)}{\boldsymbol{\beta}} = \underset{(q\times 1)}{\mathbf{c}}$$

其中,L 和 c 包含了特定的常数,且我们假设矩阵 L 是行满秩的 $q \leqslant k+1$。那么,F 统计量为:

$$F_0 = \frac{(\mathbf{Lb}-\mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{Lb}-\mathbf{c})}{qS^2}$$

如果 H_0 为真,那么,它遵循以 q 和 $n-k-1$ 为自由度的 F 统计值。

假设我们要检验一个包含两个解释变量的回归模型的"联立"零假设 $H_0:\beta_1=\beta_2=0$,我们可以取 $\mathbf{L}=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$ 和 $\mathbf{c}=[0,\ 0]'$。为了检验回归系数相等,$H_0:\beta_1=\beta_2$(等价于 $H_0:\beta_1-\beta_2=0$),我们取 $\mathbf{L}=[0,\ 1,\ -1]$ 和 $\mathbf{c}=[0]$。[61]

在下文,我们会提到在回归模型假设下,回归系数的最小二乘估计量等价于最大似然估计量。因此,当样本量够大,我们可以用 Delta 方法来推导回归系数的非线性函数的标准差。

例如,考虑如下二项式回归模型:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \qquad [4.6]$$

该模型可以用 Y 关于 x 和 x^2 的线性最小二乘法回归拟合,

因为其系数 β_0、β_1、β_2 是线性的。假设我们对回归方程达到最大值或最小值时的 x 值比较感兴趣[62]，对方程 4.6 两边取期望值，然后对 x 求导，可得：

$$\frac{dE(Y)}{dx} = \beta_1 + 2\beta_2 x$$

使之等于 0，解方程，可得到函数为最小值（假如 β_2 是正的）或者最大值时（假如 β_2 是负的）的 x 值：

$$x = \frac{-\beta_1}{2\beta_2}$$

其为回归系数 β_1、β_2 的非线性函数。

要运用 Delta 方法，我们需要 $\gamma = f(\beta_1, \beta_2) \equiv -\beta_1/(2\beta_2)$ 对回归系数求偏导数：

$$\frac{\partial\gamma}{\partial\beta_1} = \frac{-1}{2\beta_2}$$

$$\frac{\partial\gamma}{\partial\beta_2} = \frac{\beta_1}{2\beta_2^2}$$

现在，我们要计算 β_1、β_2 的最小二乘法估计 B_1、B_2 及其方差 $\hat{V}(B_1)$ 和 $\hat{V}(B_2)$，还有它们的协方差 $\hat{C}(B_1, B_2)$。我们知道，γ 的最大似然法估计为 $\hat{\gamma} = -B_1/2B_2$，那么，$\hat{\gamma}$ 的 Delta 方法方差为：

$$\hat{v}(\hat{\gamma}) = \hat{V}(B_1)\left(-\frac{1}{2B_2}\right)^2 + \hat{V}(B_2)\left(\frac{B_1}{2B_2^2}\right)^2 + 2\hat{C}(B_1, B_2)\left(-\frac{1}{2B_2}\right)\left(\frac{B_1}{2B_2^2}\right)$$

那么，γ 的 95%置信区间为 $\hat{\gamma} \pm 1.96\sqrt{\hat{v}(\hat{\gamma})}$。

第 5 节 | 回归模型的最大似然法估计

如前文所述,在线性模型假设下,$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$。所以,对于第 i 个观测,$Y_i \sim N_n(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$,其中,$\mathbf{x}_i'$ 为模型矩阵 $\mathbf{X}$ 的第 i 行。将其写成方程形式,可知第 i 个观测的概率密度为:

$$p(y_i) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{2\sigma^2}\right]$$

由于 n 个观测是独立的,因此,它们的联合概率密度为其边缘密度的乘积:

$$p(\mathbf{y}) = \frac{1}{(\sigma\sqrt{2\pi})^n}\exp\left[-\frac{\sum(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right] \qquad [4.7]$$

虽然这个方程同样遵循 $\mathbf{y}$ 的多元正态分布,但是从 $p(y_i)$ 到 $p(\mathbf{y})$ 的推导过程有助于我们考虑随机回归元。

从方程 4.7 中,我们可得对数似然函数:

$$\log_e L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \qquad [4.8]$$

为了最大化似然函数，我们需要求方程 4.8 对参数 $\boldsymbol{\beta}$, σ^2 的偏导数。当我们注意到$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$实际上是误差平方和时，求导过程会变得简单：

$$\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}'\mathbf{y})$$

$$\frac{\partial \log_e L(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

让偏导数等于 0 并解方程，可得：

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\hat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$$

最大似然法估计量 $\hat{\boldsymbol{\beta}}$ 和最小二乘法估计量 $\mathbf{b}$ 是一致的。实际上，不需要正式的最大似然法，我们也可由方程 4.7 发现这个等价关系：当负指数很小的时候，似然度会很大，且指数的分子中包含了误差平方和。因此，最小化残差平方和等价于最大化似然度。

$\hat{\sigma}^2$ 的最大似然估计量是有偏的，因此，我们会选择如前所述的类似且无偏的估计量 $S^2 = \mathbf{e}'\mathbf{e}/(n-k-1)$。然而，随着 n 的增大，$\hat{\sigma}^2$ 的偏差越来越趋近于 0，作为一个最大似然法估计量，$\hat{\sigma}^2$ 是一致的。

第6节 | 随机矩阵应用

本章中,我们进一步发展了线性回归分析理论,它不再局限于模型矩阵 $\mathbf{X}$ 是固定的这一前提。如果重复一个研究,我们希望因变量 $\mathbf{y}$ 能变化,但是由于 $\mathbf{X}$ 是固定的,那么在重复研究中,自变量的值为常数。这种情形描述了实验的真实情况,因为自变量是由研究者控制的。然而对于大多数的社会学研究来说,数据都是观测到而不是实验控制得来的。在一个观测研究中(例如,调查研究),我们一般会在重复研究中得到不同的解释变量。所以,在观测研究中,$\mathbf{X}$ 是随机而非固定的。

只要符合某些条件,线性回归统计学理论就同样适用于 $\mathbf{X}$ 是随机的情况。对于固定的自变量,其前提假设为 $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$,即所有模型矩阵的离散行的误差分布是相同的。当 $\mathbf{X}$ 为随机变量时,我们需要假设这个性质对于样本总体中所有可能的自变量组合都成立,即假设 $\mathbf{X}$ 和 $\boldsymbol{\varepsilon}$ 是独立的,那么,样本中取值为 $\varepsilon \mid \mathbf{X}_0$ 的自变量误差的条件分布为 $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$,不管选取的是哪个特定样本,$\mathbf{X}_0 = \{x_{ij}\}$。

因为 $\mathbf{X}$ 是随机的,所以它存在一些(多元)概率分布。我们不需要对这些分布给定假设,但是我们有如下要求:(1)对 $\mathbf{X}$ 的测定不存在误差,且 $\mathbf{X}$ 和 $\boldsymbol{\varepsilon}$ 是独立的(如前所述);(2)假

设 $\mathbf{X}$ 的分布与模型回归参数 $\boldsymbol{\beta}$、σ^2 无关;(3)规定 X 的协方差矩阵是非奇异的(即在总体中没有 X 是不变的,或者说没有一个 X 是其他变量的完美线性函数)。我们不用假设回归元(和误差相比较)是正态分布的,这样会好很多,因为许多 X 是非正态的,如虚拟变量和多项式变量,还有其他许多定量解释变量。[63]

虽然没必要不断重复,但是我会指出随机解释变量在新假设下的一些关键结果。其他结果可以此类推。

对于 X 值的一个特定样本 $\mathbf{X}_0$,$\mathbf{y}$ 的条件分布为:

$$\begin{aligned}E(\mathbf{y}|\mathbf{X}_0)&=E[(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})|\mathbf{X}_0]\\&=\mathbf{X}_0\boldsymbol{\beta}+E[\boldsymbol{\varepsilon}|\mathbf{X}_0]\\&=\mathbf{X}_0\boldsymbol{\beta}\end{aligned}$$

那么,最小二乘法估计量的条件期望均值为:

$$\begin{aligned}E(\mathbf{b}|\mathbf{X}_0)&=E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}_0]\\&=(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'E[\mathbf{y}|\mathbf{X}_0]\\&=(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{X}_0\boldsymbol{\beta}\\&=\boldsymbol{\beta}\end{aligned}$$

因为这个过程可以对任意 $\mathbf{X}$ 进行重复,所以最小二乘估计量 $\mathbf{b}$ 对于任意该类值都是条件无偏的,它在无条件下也是无偏的,$E(\mathbf{b})=\boldsymbol{\beta}$。

现在我们对 $\boldsymbol{\beta}$ 进行统计估计。具体来说,想象我们需要计算联合零假设 $H_0:\beta_1=\cdots=\beta_k=0$ 的 p 值。因为当 $\mathbf{X}$ 为固定的时候,$\boldsymbol{\varepsilon}\mid\mathbf{X}_0\sim N_n(\mathbf{0},\sigma^2\mathbf{I}_n)$,对于 $\mathbf{X}=\mathbf{X}_0$ 的 p 值是正

确的(即对于正在使用的样本)。然而,$\mathbf{X}_0$ 没有什么特别之处,误差向量 $\boldsymbol{\varepsilon}$ 是独立于 $\mathbf{X}$ 的,所以对于任意的 $\mathbf{X}$, $\boldsymbol{\varepsilon}$ 的分布都为 $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ 。因此,p 值是无条件有效的。

最后,我要指出,$\boldsymbol{\beta}$、σ^2 的最大似然估计量不会因 $\mathbf{X}$ 是随机的而改变,只要新的假设成立——当 $\mathbf{X}$ 为随机变量时,样本观测不仅包含因变量$(Y_1, \cdots, Y_n)$,而且包含自变量$(\mathbf{x}_1', \cdots, \mathbf{x}_n')$;我们可以把观测记为$[Y_1, \mathbf{x}_1'], \cdots, [Y_n, \mathbf{x}_n']$。由于这些观测是独立采样的,因此,它们的概率密度为它们的边缘密度的乘积:

$$
\begin{aligned}
p(\mathbf{y}, \mathbf{X}) &\equiv p([y_1, \mathbf{x}_1'], \cdots, [y_n, \mathbf{x}_n']) \\
&= p(y_1, \mathbf{x}_1') \times \cdots \times p(y_n, \mathbf{x}_n')
\end{aligned}
$$

第 i 次观测的概率密度 $p(y_i, \mathbf{x}_i')$,可以写成 $p(y_i | \mathbf{x}_i')p(\mathbf{x}_i')$。根据线性模型,给定 $\mathbf{x}_i'$的 y_i 的条件分布是正态的:

$$
p(y_i \mid \mathbf{x}_i') = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right]
$$

那么,所有观测的联合概率密度为:

$$
\begin{aligned}
p(\mathbf{y}, \mathbf{X}) &= \prod_{i=1}^{n} p(\mathbf{x}'_i)\,\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(y_i - \mathbf{x}'_i\boldsymbol{\beta})^2}{2\sigma^2}\right] \\
&= \left[\prod_{i=1}^{n} p(\mathbf{x}'_i)\right]\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right] \\
&= p(\mathbf{X})p(\mathbf{y} \mid \mathbf{X})
\end{aligned}
$$

只要 $p(\mathbf{X})$不再由参数 $\boldsymbol{\beta}$、σ^2 决定,我们就可以在最大化 $p(\mathbf{y}, \mathbf{X})$的过程中,忽略 X 的联合密度。最后,对于固定的 $\mathbf{X}$,最大似然估计量 $\boldsymbol{\beta}$ 与最小二乘估计量一致。

注释

[1] 我们可以为长方形矩阵定义一个广义的逆矩阵，但是对于方形矩阵，其逆矩阵不合常规。

[2] det **A** 的另一种常用的替换表示法为$|\mathbf{A}|$。

[3] 在统计问题中应用几何向量，尽管在向量空间的维数通常与样本量 n 相等，我们仍可以根据我们的兴趣将子空间限制到二维或者三维。

[4] 矩阵 $\mathbf{X}'\mathbf{X}$ 的元为 $\mathbf{x}_i'\mathbf{x}_j=\mathbf{x}_i\cdot\mathbf{x}_j$，其中，$\mathbf{x}_i$ 和 $\mathbf{x}_j$ 分别为矩阵 $\mathbf{X}$ 的第 i 列和第 j 列。类似地，$\mathbf{X}'\mathbf{X}$ 的第 i个对角元为 $\mathbf{x}_i'\mathbf{x}_i=\mathbf{x}_i\cdot\mathbf{x}_i$。

[5] 按照常规，我们通常定义较小的角为两向量间的夹角(因此，该角度不可能大于 180°)，记做 w，那么较大的角则为 $360-w$。由于 $\cos(360-w)=\cos(w)$，因此，这样定义不会引起歧义。

[6] 不要把线性方程的几何表示与向量的几何表示混淆。

[7] 有关术语的说明：一些作者不论方程组一致与否，或者系数矩阵的秩多少，一概把“方程数目大于未知数”的方程组定义为“超定方程组”，把“方程数目小于未知数”的方程组定义为“欠定方程组”。我认为，我在文中的论述(Daivs, 1973)更合适些。

[8] 有关广义逆矩阵在统计学中的延伸论述，请参考相关著作(Rao & Mitra, 1971)。

[9] 我们可以给方程 1.12 加一些限制条件，使得广义逆矩阵变得唯一。比如，Moore-Penrose 广义逆矩阵 $\mathbf{A}^+$ 满足四个条件：$\mathbf{A}\mathbf{A}^+\mathbf{A}=\mathbf{A}$；$\mathbf{A}^+\mathbf{A}\mathbf{A}^+=\mathbf{A}^+$；$\mathbf{A}\mathbf{A}^+$ 是对称的；$\mathbf{A}^+\mathbf{A}$ 也是对称的。在典型的统计应用中，广义逆矩阵非常好用。

[10] 首先，$\mathbf{A}_C$ 是 $\mathbf{A}_C'$的广义逆矩阵；其次，由方程 1.15，我们知道 $\mathbf{A}=\mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1}$，那么，

$$
\begin{aligned}
\mathbf{A}\mathbf{A}^-\mathbf{A} &= (\mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1})(\mathbf{E}^*\mathbf{A}_C'\mathbf{E})(\mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1})\\
&= \mathbf{E}^{-1}\mathbf{A}_C\mathbf{A}_C'\mathbf{A}_C\mathbf{E}^{*-1}\\
&= \mathbf{E}^{-1}\mathbf{A}_C\mathbf{E}^{*-1}\\
&= \mathbf{A}
\end{aligned}
$$

从而方程成立。

[11] 一元二次方程回顾：x 满足方程：

$$ax^2+bx+c=0$$

其中，a、b 和 c 为指定常数，那么，

$$x=\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

[12] 通过解特征方程求特征值不是一个很好且很有吸引力的方法，还有一些比较实际的求解特征值及其特征向量的方法。

[13] 对于一个对称正定矩阵，我们仍然有可能找到其 Cholesky 因子，但是加上相应行中的其他元素，矩阵 **U** 会有一个或多个对角元素为 0。另外，为了解决矩阵 **U** 对角元素的问题，我们必须取正平方根。

[14] 牛顿声称，莱布尼茨窃用了他的研究成果，从而引发了科学史上最著名的争论之一。

[15] 在有些数学领域，自然数仅指正整数。

[16] 虽然我倾向于详细地标出对数的底，如 $\log_{10}$，$\log_e$(除非底是不相关的，标出 log 已经足够)，但是其他许多学者喜欢用 log 或者 ln 代替自然对数。

[17] 微商不是普通的数字，所以可以把链式法则想象成同时乘以和除以一个引发机制微商 dz。在导数中引入微商是有效的。

[18] 超平面是指超出三维空间的一个线性(平的)表面。超平面的维度比总空间的维度少 1，就像三维空间里嵌入的一个二维对象。

[19] 有些学者喜欢加入约束条件，而不是减去之：

$$h(x_1, x_2, \cdots, x_n, \lambda)\equiv f(x_1, x_2, \cdots, x_n)+\lambda\times g(x_1, x_2, \cdots, x_n)$$

但是，除了 λ 的符号发生改变以外，这两种方法是没有差异的。

[20] 如果对于任意非零向量 **x**，有 $\mathbf{X}'\mathbf{H}\mathbf{x}>0$，那么我们说方阵 **H**(这里是海森矩阵)是正定的。正定海森矩阵是最小值的充分非必要条件。同样，如果对于任意非零向量 **x**，有 $\mathbf{X}'\mathbf{H}\mathbf{X}<0$，那么我们说方阵 **H**(这里是海森矩阵)是负定的。负定海森矩阵是最大值的充分非必要条件。

[21] 非零整数 n 的阶乘定义为 $n!=n(n-1)(n-2)\cdots(2)(1)$。按照惯例，$0!=1!=1$。

[22] 这种近似叫做“穷举法”(虽然不是传统的极限表示)，被古代希腊人所熟悉。

[23] 读者可以证明 $F(x)$ 是函数 $f(x)=x^2+3$ 的一个反导数。一般而言，我们可以通过反过来应用幂函数的导数的规则，进而寻找多元函数的反导数。

[24] 样本空间是无限的，原因在于可能需要等任意足够长的时间才能观测到第一次出现硬币正面的情况，尽管在现实中等待无限长时间的概率是极小的。通常，当 S 是离散且无限的时候，我们称其为“可数无穷”，因为 S 中的元素与自然数 0、1、2 等一一对应。

[25] 这些定理类似于(或者等价于)那些由 20 世纪俄国数学家 A. N. 柯尔莫戈洛夫(A. N. Kolmogorov)所提出的定理。

[26] 如果随机变量 X 可取有限或者可数无穷多的不同值,那么我们说,该随机变量 X 是离散的。

[27] 概率通常对应的是密度函数 $p(x)$下的区域,随机变量 X 的某一特殊值 x_0 下的区域——一条与横坐标垂直的线,其概率为 0。

[28] 一个连续随机变量 X 的概率密度函数常常表示为 $f(x)$,其累积分布函数表示为 $F(x)$,但是我觉得用 $p(x)$和 $p(x)$表示更好,因此我倾向于把 $f(\cdot)$留作他用,比如随机变量的变换。

[29] 如果你对积分不是十分熟悉,不要过于苛求。其理解的关键点在于,我们将概率密度曲线 $p(x)$以下的区域解释为概率,累积分布函数 CDF $P(x)$的高度告诉我们,随机变量 X 的可观测值小于或等于某一特殊值 x 的概率。积分符号 $\int$ 表示连续求和,代表了曲线以下的区域。

[30] 由于 $p(x)$具有连续性,再加上 $\Pr(X = x_0) = \Pr(X = x_1) = 0$,我们不需要区别 $\Pr(x_0 \leqslant X \leqslant x_1)$ 和 $\Pr(x_0 < X < x_1)$。

[31] 一些随机变量没有定义其期望值和方差,在这里,我忽略了这一可能性。

[32] 我们是在 X 的支持下整合的,因此不需要包含整条实线。

[33] 如果你对第 2 章介绍的有关积分部分的内容不太熟悉,其实可以简单地把积分符号 $\int$ 看做求和符号 $\sum$ 。

[34] 这里用希腊字母 π 的原因在于,概率无法被直接观测到。由于 π 代表概率,其值在 0 到 1 之间,所以不要将其跟数学常数 ≈ 3.1416 混淆。

[35] 回忆一下有关阶乘运算的法则:

$$n! \equiv n \times (n-1) \times \cdots \times 2 \times 1 \ (n \text{ 为任意大于 1 的整数})$$

$$\equiv 1 (n \text{ 等于 0 或 1})$$

[36] 一些作者会用 $N(\mu, \sigma)$来代表正态分布,该表达用正态分布的标准方差代替了我们所用的方差。

[37] 任意存在有限均值和方差的随机变量都可以标准化为均值为 0、方差为 1 的随机变量。但是标准化对分布的形状并无影响,尤其是,它不会把一个不是正态分布的变量变为正态分布的变量。

[38] 小写字母 t 是一个通用表示方法。

[39] 当 $n = 1$ 时,$E(t)$的期望值不存在,但是 t 的中位数和众数仍为 0。t_1 被称为"柯西分布"(Cauchy distribution),它是以 19 世纪法国数学家奥古斯丁·路易斯·柯西的名字命名的。

[40] 将命题反过来则为假:随机向量 $\mathbf{x}$ 所包含的元素的边缘分布是单因素

正态分布，不一定为多元正态分布。

[41] 把$\{p_1, p_2, \cdots, p_n, \cdots\}$说成非随机序列不会有明显的矛盾，虽然这些概率建立在随机变量上，但是概率本身是特定的数字——如 0.6、0.9 等等。

[42] 用渐近分布定义渐近期望值和方差更具吸引力，因为这个目的的序列不是在所有情况下都存在的（Theil, 1971: 375—376; McCallum, 1973）。我所用的渐近期望和方差的符号——$\varepsilon(\cdot)$和$\upsilon(\cdot)$——不是标准化的，读者应该注意，这些符号有时会被普通期望值和方差的符号——$E(\cdot)$和$V(\cdot)$——所替代。

[43] 有关统计估计的大部分材料以及这本书的相关内容均来自费舍尔（Fisher, 1992）的一篇论文，该论文被誉为 20 世纪最重要的统计论文之一（Aldrich, 1997）。

[44] 如果没有强调对称性，我们所说的中心概念就会变得很模糊。

[45] 严格来说，ρ_{LAV}的导数没有定义 $E=0$ 时的情况，但为方便起见，我们将 $E=0$ 的情况设定为 $\psi_{\mathrm{LAV}}(0)\equiv 0$。

[46] 可以写成该形式的估计值可以看成最大似然估计值的广义形式，因此，也被称为“M 估计值”。最大似然估计量是通过对合适的概率函数或者概率密度函数 $p(\cdot)$进行变换 $\rho_{\mathrm{ML}}(x-\mu)\equiv -\log_e p(x-\mu)$ 得来的。

[47] 我所用的命名不太严格，只是比较方便而已。严格地说，ψ 函数不是一个影响函数，只是其形状与影响函数相同。

[48] “双平方”常常应用在 ψ 函数和权重函数上（因此称为“双权”），它是近期才出现的统计词汇，但是作为目标函数却不是。

[49] 因为一个重新降级过的 M 估计值的估计方程（如双平方）可以有多于一个的平方根，所以选择初始估计便成为必然。

[50] 似然函数是 π 取值在 0 和 1 之间的连续函数。此例与概率函数的不同之处是，它所有可能的样本是有限的，为 2^{10}。

[51] 严格地说，Fisher 信息是参数值为 α 时所估计记分的方差：

$$I(\alpha)=E\left[\left(\frac{d\log_e L(\alpha)}{d\alpha}\right)^2 \middle| \alpha\right]$$

在许多情况下，该方程与文中提及的方程等价，但是相对复杂和麻烦。请注意，记分的方差仅是其平方的期望值，因为在 α 时，其期望记分为 0。

[52] 在观测信息上建立 MLE$\hat{\alpha}$ 的方差估计值是可能的，而且更方便计算。

$$I_O(\hat{\alpha})\equiv\frac{d^2\log_e L(\hat{\alpha})}{d\hat{\alpha}^2}$$

[53] 我们说参数是独立的，意思是空值可以从其他参数取值得到。如果参数间存在依附关系，那么，多余的参数就会通过一个函数被其他参数所替代。

[54] 在单一参数的例子中，Fisher 信息更广义的定义为：

$$I(\boldsymbol{\alpha})=E\left[\left(\frac{\partial \log_e L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right)^2 \middle| \alpha\right]$$

同样，我们可以通过用 MLE$\hat{\alpha}$ 时的观测信息来进行研究。

[55] 要利用贝叶斯推断，先验想法必须符合概率论，所有先验概率的和必须为 1。

[56] 在这种情况下，先验是一个长方形密度函数，其参数 π 被限制在 0 到 1 之间。例如，估计正态分布的均值 μ 不存在界域问题，那么它在 $-\infty<\mu<\infty$ 的扁平先验形式 $p(\mu)=c$ 不存在一个有限概率，因此无法代表密度函数。当它和似然函数合并后，例如成为一个不正常先验，还是会导致一个正常后验分布——一个积分为 1 的后验分布。还有，一个概率模型的参数化扁平先验对于一个可替换参数化方案来说并不扁平。假设我们取发生比 $\omega\equiv\pi/(1-\pi)$ 作为参数，其等价的对数形式为 $\lambda\equiv\log_e[\pi/(1-\pi)]$。$\pi$ 的扁平先验对 ω 和 λ 都是不扁平的。

[57] 一个替代方法是把后验分布的均值作为 π 的点估计。然而在大多数情况下，随着样本量的不断增加，后验分布会越来越趋近正态分布，那么，如果样本量足够大，其众数和均值是几乎相等的。

[58] 该规则有个例外，即对于某些参数值，先验分布为零密度分布，那么，对于这些参数值，其后验分布也会是零密度分布。

[59] 一般而言，独立意味着不相关，但是不相关的随机变量不一定是独立的。但是，在多元正态分布中，独立和不相关是等价的。

[60] 关于这段所提到的结果的证明，可参见相关著作(Rao, 1973)。

[61] 为了使假设可行，两个解释变量 x_1、x_2 必须用相同的单位来衡量。

[62] Delta 方法在这个问题上的应用是由韦斯伯格(Weisberg, 2005)提议的。

[63] 前面的结果对于新假设下的 X 是适用的，但并不代表这些假设是完全没有问题的。许多解释变量都是有测量误差的，且在某些条件下，它们会使估计系数有严重的偏差。同样，在特定的(一般的)回归等式理解中，关于解释变量误差独立的假设等价于模型中包含的 Y 的决定因素和忽略因素是不相关的。最后，线性假设、常误差方差和正态都是有潜在问题的。能令人满意地处理好这些问题，与回归分析作为数学抽象和数据分析的实用工具是不一样的。

参考文献

Aldrich, J. (1997). R. A. "Fisher and the making of maximum-likelihood 1912—1922." *Statistical Science*, *12*, 162—176.

Binomore, K., & Davies, J. (2001). *Calculus: Concepts and methods*. Cambridge, UK: Cambridge University Press.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Davis, P. J. (1973). *The mathematics of matrices: A first book of matrix theory and linear algebra* (2nd ed.). Lexington, MA: Xerox College.

Engle, R. F. (1984). "Wald, likelihood ratio, and Lagrange multiplier tests in economics." In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of economics* (Vol.2, 775—879) Amsterdam: North-Holland.

Fisher, R. A. (1992). "On the mathematical foundations of theoretical statistics." *Philosophical transactions of the Royal Society of London* A. 222, 309—368.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Graybill, F. A. (1983). *Introduction to matrices with applications in statistics* (2nd ed.). Belmont, CA: Wadsworth.

Green, P. E., & Carroll, J. D. (1976). *Mathematical tools for applied multivariate analysis*. New York: Academic Press.

Healy, M. J. R. (1986). *Matrices for statistics*. Oxford, UK: Clarendon Press.

Johnston, J. (1972). *Econometric methods* (2nd ed.). New York: McGraw-Hill.

Kennedy, W. J., Jr., & Gentle, J. E. (1980). *Statistical computing*. New York: Dekker.

Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Oxford, UK: Blackwell.

McCallum, B. T. (1973). "A note concerning asymptotic covariance expressions." *Econometrica*, *41*, 581—583.

Monahan, J. F. (2001). *Numerical methods of statistics*. Cambridge, UK: Cambridge University Press.

Namboodiri, K. (1984). *Matrix algebra: An introduction*. Beverly Hills,

CA: Sage.

Rao, C. R. (1973). *Linear Statistical inference and its applications* (2nd ed.). New York: Wiley.

Rao, C. R., & Mitra, S. K. (1971). *Generalized inverse of matrices and its applications*. New York: Wiley.

Searle, S. R. (1982). *Matrix algebra useful for statistics*. New York: Wiley.

Theil, H. (1971). *Principles of econometrics*. New York: Wiley.

Thompson, S. P., & Gardner, M. (1998). *Calculus made easy*. New York: St. Martin's Press.

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). New York: Wiley.

Wonnacott, T. H., & Wonnacott, R. J. (1990). *Introductory statistics* (5th ed.). New York: Wiley.

Zellner, A. (1983). "Statistical theory and econometrics." In Z. Griliches & M. D. Intriligator(Eds.), *Handbook of econometrics* (Vol. 1, 67—178). Amsterdam: North-Holland.

译名对照表

asymptotic bias	渐近偏差
asymptotic distribution theory	渐近分布理论
augmented matrix	增广矩阵
basis	基
Bayesian statistical inference	贝叶斯统计推断
Bernoulli distribution	伯努利分布
binomial distribution	二项分布
biweight	双权
bisquare	双平方
bounded	有界的
breakdown point	崩溃点
canonical form	标准形式
central-limit theorem	中心极限定理
central postprior interval	中央后验区间
characteristic equation	特征方程
characteristic root	特征根
chain rule	链式规则
comfortable for multification	乘法相适
complement	补集
conditional probability	条件概率
conditional probability density	条件概率密度
conjugate priors	共轭先验
definite integral	定积分
derivative	导数
diagonal matrix	对角矩阵
difference quotient	差商
differentiation	微分
efficiency	有效性
eigenvalue	特征值
eigenvector	特征向量
empty event	空事件

entry	元
expected information	期望信息
extrema	极值
factorial	阶乘
Fisher information	Fisher 信息
fitted value	拟合值
fractional powers	分数幂函数
Gaussian elimination	高斯消去法
hazard	风险
Hessian matrix	海森矩阵
homogeneous system of equations	齐次方程组
hyperplane	超平面
idempotent	等幂元
identity matrix	单位矩阵
influence function	影响函数
inner product	内积
integer	整数
irrational number	无理数
Jacobian of the transformation	雅可比迭代
joint probability distribution	联合概率分布
jointly sufficient statistics	联合充分统计量
Lagrange-multiplier test	拉格朗日乘数检验
latent growth curve model	潜在增长曲线模型
latent root	潜伏根
least-absolute values	最小绝对值
limit	极限
limit of integration	积分域
linear simultaneous equation	线性联立方程
log-odds	对数优比
lower-triangular matrix	下三角矩阵
marginal posterior distribution	边缘后验分布
marginal probability distribution	边缘概率分布

marginal probability density	边缘概率密度
matrices of cross-products	交叉乘积矩阵
matrices of sums of squares	平方和矩阵
matrix inverse	逆矩阵
maximum-likelihood estimation	最大似然估计
mean-deviation vector	平均偏差向量
mean-squared error	均方误差
minimally sufficient	最低充分
multinomial distribution	多项分布
multivariate-normal distribution	多元正态分布
natural number	自然数
negative binomial distribution	负二项分布
negative powers	负幂函数
non-singular matrix	非奇异矩阵
nonstochastic infinite sequence	非随机无限序列
nontrivial solution	非平凡解
null event	零事件
objective function	目标函数
odds	发生比
operator	算子
optimization	最优化
order	阶
orthogonality	正交
orthogonal projection	正交投影
outlier	异常值
overdetermined system of equations	超定方程组
partial regression coefficient	偏回归系数
particular value	表示变量的特殊值
partitioned matrix	分块矩阵
pivot	主元
positive-definite matrices	正定矩阵
point of inflection	拐点

Poisson distribution	泊松分布
posterior probability	后验概率
prior probability	先验概率
pythagorean theorem	勾股定理
quotient	差商
random variable	随机变量
rank	秩
rational number	有理数
real number	实数
realization	实现
rectangular/uniform distribution	均匀分布
reduced row-echelon form	行简化阶梯形矩阵
regression sum of squares	回归平方和
residual sum of squares	残差平方和
resistance	耐抗性
robust regression	稳健回归
robustness of validity	效度稳健性
sample space	样本空间
sampling distribution	抽样分布
sampling variance	抽样方差
scalar constant	纯量常数
scalar random variable	纯量随机变量
scale parameter	尺度参数
score	记分
secant	割线
shape parameter	形状参数
singular-value decomposition	奇异值分解
singular matrix	奇异矩阵
stationary point	驻点
submatrix	子矩阵
support	支持
tangent line	切线

tangent hyperplane	切超平面
tangent plane	切面
trace	迹
transpose	转置
total sum of squares	总平方和
tuning constant	细调常数
unbiased	无偏
unconditional probability	无条件概率
undetermined system of equations	欠定方程组
unit-normal distribution	单位正态分布
upper-triangular matrix	上三角矩阵
vector	向量
vector random variable	随机向量
vector partial derivative	向量偏导数